Universidad del Desarrollo
Facultad de Ingeniería

"PREDICCIÓN DE LA INFLAMABILIDAD DE PRODUCTOS QUÍMICOS DEL DATASET CRAMER-UDD MEDIANTE MACHINE LEARNING"

POR: SOFÍA VITS CONTRERAS

Proyecto de grado presentado a la Facultad de Ingeniería de la Universidad del Desarrollo para optar al grado académico de Magíster en Data Science

PROFESOR GUÍA: Dr. GERMÁN GÓMEZ
CO-TUTOR: Dr. JUAN ELIZALDE

Diciembre 2025
SANTIAGO

A mi abuela Marta, quien graciosamente combinó el arte y la lógica.

AGRADECIMIENTO

Al Dr. Juan Elizalde, por haber suministrado los datasets de fórmulas e ingredientes de la empresa Cramer Productos Aromáticos S.A.C.I. y por haber guiado este proyecto desde una perspectiva química.

TABLA DE CONTENIDO

Resumen

Los productos químicos son clasificados en categorías de riesgo de acuerdo a su reactividad, inflamabilidad, perjuicios a la salud, y otros riesgos especiales relacionados con la naturaleza fisicoquímica de una fórmula. En este trabajo se analizará específicamente el punto de inflamación de una fórmula química. Esta característica es uno de los parámetros utilizados para evaluar la inflamabilidad de una sustancia química.

La determinación de esta propiedad requiere la aplicación de mediciones experimentales específicas, las cuales pueden resultar ser muy costosas para una empresa desde los puntos de vista económico y temporal. Por este motivo, la empresa Cramer compartió sus datos para poder crear modelos de clasificación multiclase, con el propósito de clasificar la inflamabilidad de sus fórmulas líquidas. Las fórmulas fueron clasificadas en fórmulas que contienen compuestos líquidos con comportamiento gaseoso (más específicamente compuestos azufrados), fórmulas etanólicas, fórmulas acuosas, y fórmulas que contienen otros solventes orgánicos.

Los modelos aplicados fueron regresión logística multiclase, Random Forest, LightGBM, y CatBoost. Todos estos modelos fueron calibrados mediante predicción conformal utilizando *margin nonconformity score* para dicho propósito. Posteriormente se realizó un análisis SHAP de los modelos LightGBM y CatBoost, comparando la importancia asignadas a cada variable por estos modelos, así como el estudio de los gráficos de dependencia de dichas variables.

1. Introducción

Las empresas de productos químicos sintetizan, mezclan, almacenan, y comercializan una gran cantidad de compuestos potencialmente peligrosos, los cuales deben ser sometidos a experimentos para determinar su clasificación de riesgo. Un tipo de compuesto peligroso son los productos inflamables, los cuales poseen diversas propiedades que definen su perfil de riesgo.

Una de las características relacionadas con la clasificación de riesgo de los compuestos inflamables es el punto de inflamación, el cual cobra especial relevancia al fabricar, almacenar, transportar, y vender compuestos orgánicos volátiles. En este Capstone Project me enfocaré en predecir la inflamabilidad de productos químicos creados por la empresa Cramer utilizando el punto de inflamación de diversos ingredientes que componen sus fórmulas.

Esta empresa nacional está enfocada en el mercado latinoamericano, la cual ha fabricado más de 28.000 fórmulas de saborizantes y fragancias en estado líquido para uso industrial. Las fórmulas distribuidas por esta empresa contienen un amplio rango de componentes, lo cual implica una dificultad mayor para predecir su punto de inflamación. La mayoría de los métodos publicados al respecto en revistas científicas se enfocan en mezclas binarias, ternarias, y en algunos casos analizan mezclas que contienen aproximadamente seis ingredientes en total.

Muchos de estos componentes son compuestos orgánicos volátiles, los cuales pueden ser líquidos y vapores extrememadamente inflamables, muy inflamables o

inflamables. Por este motivo es necesario contar con normativa que regule su fabricación, almacenamiento, transporte y comercialización.

Una de las normas chilenas relacionadas con el tratamiento de sustancias peligrosas es el Decreto 57 del Ministerio de Salud exige determinar el punto de inflamación de las mezclas fabricadas y comercializadas por Cramer, considerando los requisitos del Sistema Globalmente Armonizado de Clasificación y Etiquetado de Productos Químicos (GHS) para su determinación. Existe el método de copa cerrada para determinar dicho parámetro, el cual es un procedimiento manual, lento y requiere un equipo que cuesta una alta suma de dinero. Este equipo debe ser sometido a mantención preventiva, y debe ser operado por personal acreditado en la empresa. Además, las empresas compradoras solicitan una entrega expedita de los productos encargados. Para lograr esto muchas veces es necesario crear varios prototipos antes de obtener la mezcla adecuada para el cliente, los cuales requieren a su vez la determinación de su punto de inflamación a pesar de no ser comercializados posteriormente.

Por lo tanto, se requiere implementar modelos de machine learning que permitan facilitar la clasificación de estas fórmulas de acuerdo a su punto de inflamación, agilizando los procesos logísticos de la empresa Cramer.

2. Trabajo Relacionado

El punto de inflamación se define como la temperatura más baja a la que una mezcla de una sustancia con el aire exterior se inflama y arde inmediatamente al entrar en contacto con una llama. Es una propiedad física de los compuestos orgánicos que resulta crucial en aplicaciones industriales y experimentos de laboratorio, dado que proporciona información sobre las precauciones necesarias para manipular diversos materiales orgánicos. Esta propiedad se considera un factor crítico para clasificar la inflamabilidad de líquidos (Amirkhani et al, 2022; Cao et al, 2020).

Los líquidos inflamables se clasifican en tres categorías: categoría 1 (punto de inflamación inferior a 23°C y punto inicial de ebullición igual o inferior a 35°C), categoría 2 (punto de inflamación inferior a 23°C y punto inicial de ebullición superior a 35°C), y categoría 3 (punto de inflamación igual o superior a 23°C e igual o inferior a 60°C). Las indicaciones de peligro para cada categoría son líquidos y vapores extremadamente inflamables, líquidos y vapores muy inflamables, y líquidos y vapores inflamables (MINSAL, 2021).

Experimentalmente, se aplican pruebas de copa abierta y copa cerrada para determinar los puntos de inflamación de las mezclas. Estos métodos son costosos y lentos, por lo cual se han desarrollado diversos métodos para predecir el punto de inflamación de sustancias puras y mezclas. Para ello, existen varios enfoques posibles, entre los que se incluyen la regresión empírica de datos, los métodos basados en la presión de vapor y el método de relación estructura cuantitativa-propiedad (QSPR).

Los métodos basados en la presión de vapor suelen emplear la regla de Le Chatelier y el cálculo de equilibrios vapor-líquido (VLE) para la predicción del punto de inflamación. Los métodos de este tipo son Margules de tres sufijos, NRTL, Wilson, UNIQUAC, y UNIFAC. Esto implica resolver ecuaciones matemáticas complejas para determinar el punto de inflamación de mezclas, y pueden ser no aplicables a mezclas con más de tres componentes (Costa do Nascimento et al, 2024).

El método de relación estructura cuantitativa-propiedad (QSPR) utiliza parámetros basados en moléculas, conocidos como descriptores moleculares, para predecir el punto de inflamación de una sustancia (Cao et al, 2020; Pan et al, 2020). Mediante el análisis estadístico de los diversos parámetros estructurales y los datos experimentales de las propiedades de las moléculas, se puede establecer una relación cuantitativa entre los parámetros estructurales y las propiedades del compuesto a partir de los experimentos. Estos métodos pueden ser combinados con algoritmos de machine learning combinados con un algoritmo genético (GA-MLR, GA-SVM) y de redes neuronales para predecir el punto de inflamación. Sin embargo, se requiere la aplicación de métodos de mecánica molecular o mecánica cuántica para determinar los parámetros estructurales, lo cual implica el uso de software especializado y tiempos de cálculo que pueden ser muy extensos (Pan et al, 2020).

Los modelos empíricos se basan principalmente en el análisis de regresión de datos, que suele correlacionar el punto de inflamación de la mezcla con la entalpía de vaporización, la concentración de algún componente, el tamaño de la cadena de carbono, la presión de vapor, el punto de inflamación de los componentes puros y las temperaturas

normales de ebullición, entre otras propiedades. Algunos métodos empleados son la regresión multilineal (MLR), la regresión no lineal múltiple (MNR), modelos de redes neuronales artificiales de tipo perceptrón multicapa (MLP), sistemas de inferencia neurodifusa adaptativa (ANFIS) combinados con un algoritmo genético (GA) (Amhirkani et al, 2022). El problema es que estos modelos sólo son aplicables de forma limitada, debido a que no generalizan de forma adecuada ya que están basados en datos empíricos.

La predicción del punto de inflamación se vuelve más complicada cuando hay dos o más sustancias involucradas en una solución debido a la complejidad que conlleva un mayor número de componentes en el mismo sistema, ya que las interacciones entre diferentes especies químicas en solución tienden a diferir en intensidad y naturaleza (repulsión o atracción) de las interacciones entre sustancias puras. Siempre que se altera la proporción de una mezcla, o se cambia el tipo o la naturaleza de la mezcla, su punto de inflamación también cambia, debido a la prevalencia de fuerzas de repulsión o de atracción entre las moléculas en solución.

Otro punto a considerar es la determinación de la incertidumbre de las predicciones del valor del punto de inflamación, para lo cual se pueden utilizar métodos de remuestreo tales como bootstrap, aplicación de redes neuronales bayesianas (BNN), extensiones bayesianas a modelos como Support Vector Machines (SVM), aplicación de procesos Gaussianos (GP) (Mowbray et al, 2022), predicción conformal aplicada a modelos de machine learning y redes neuronales para obtener un intervalo de predicción que contiene al valor predicho en el centro de este intervalo y que además no asume que los datos sean independientes e idénticamente distribuidos (Jovic, 2024; Manokhin, 2025).

Finalmente, es importante interpretar los modelos de machine learning utilizados para predecir el punto de inflamación. Uno de los métodos de interpretabilidad del aprendizaje automático basados en características es el análisis SHapley Additive exPlanations (SHAP). El análisis SHAP forma parte de la fase de interpretación del modelo en un flujo de trabajo de machine learning y solo debe realizarse si el modelo demuestra un rendimiento adecuado. Por lo tanto, los valores SHAP comunicados suelen ser los valores SHAP que corresponden al conjunto de datos de prueba (Ponce-Bobadilla et al, 2024).

3. Hipótesis y Objetivos

Es posible determinar la clase de inflamabilidad de una fórmula a partir de variables derivadas del punto de inflamación de sus componentes respectivos y de su composición porcentual.

Objetivo general

Clasificar las fórmulas fabricadas y comercializadas por la empresa Cramer en las categorías no inflamables, inflamables y muy inflamables.

Objetivos específicos

1. Entrenar modelos de clasificación multiclase para determinar las clases de inflamabilidad de las fórmulas estudiadas
2. Determinar los errores críticos de clasificación de los modelos de machine learning aplicados a los grupos de fórmulas respectivas
3. Interpretar modelos de clasificación multiclase que tengan un buen rendimiento para determinar la influencia de las variables estudiadas sobre los resultados obtenidos

4. Datos y Metodología

4.1. Datos

La empresa Cramer proporcionó dos datasets para determinar la inflamabilidad de sus fórmulas. Para el caso de los ingredientes el dataset original contiene 1911 filas, y dos columnas denominadas como SUN MP (correspondiente al código de un ingrediente) y Flash Point MP, la cual contiene los puntos de inflamación de cada ingrediente.

La variable objetivo, denominada inflamabilidad, fue derivada a partir de la variable Flash Point PT, considerando las clases muy inflamable (flash point inferior a 23°C), inflamable (correspondiente a un flash point que es igual o mayor a 23°C e igual o inferior a 60°C), y la clase no inflamable que incluye a todos los valores de flash point mayores a 60°C.

Los valores de flash point consignados como NA en el dataset de ingredientes fueron imputados con la temperatura 150°C, debido a que corresponden a compuestos no inflamables. Se agregó la columna FP_NA_Flag para identificar a las columnas que contienen valores de flash point imputados.

El dataset correspondiente a las fórmulas contiene 1.096.336 filas luego de eliminar las filas duplicadas. Este dataset contiene cuatro columnas, las cuales son SUN PT, correspondiente al código de la fórmula, Flash Point PT, el cual es el flash point de la fórmula, SUN MP (columna en común con el dataset de ingredientes) y % MP en fórmula, la cual menciona el porcentaje de cada ingrediente de una fórmula específica.

Estos datos son insuficientes para clasificar correctamente a las fórmulas, por lo tanto, fue necesario realizar un proceso de feature engineering para derivar variables diferentes. Estas variables permiten que los modelos capturen la física de la mezcla y el perfil de riesgo inherente a la composición de la fórmula.

A. Métrica principal (weighted average)

Esta es la variable más importante, ya que es la aproximación lineal del punto de inflamación de la mezcla.

Variable	Significado
FP_weighted_mean	Promedio ponderado del flash point de los ingredientes. Captura la contribución directa de cada componente.

B. Estadísticas de dispersión y extremos

Estas variables describen la variabilidad y los ingredientes más extremos presentes en la fórmula.

Variable	Significado
FP_min	Flash point mínimo de todos los ingredientes. Indica el riesgo del componente más volátil.
FP_max	Flash point máximo de todos los ingredientes.
FP_std	Desviación estándar del flash point de los ingredientes.
FP_range	Rango de flash point.
FP_weighted_std	Desviación estándar ponderada por concentración.

C. Características de composición y diversidad

Estas variables miden la complejidad de la fórmula y la concentración del ingrediente dominante.

Variable	Significado
num_ingredientes	Cantidad total de ingredientes en la fórmula.
max_concentracion	Concentración del ingrediente mayoritario.
min_concentracion	Concentración del ingrediente minoritario.
concentracion_dominante	Idéntico a max_concentracion.
entropia	Entropía de Shannon (Camesasca et al, 2006) de las concentraciones. Mide cuán uniformemente están distribuidas las concentraciones. Un valor alto significa mayor diversidad.
FP_ingrediente_dominante	Flash point del ingrediente que tiene la concentración más alta.

D. Proporciones y conteos por rango de riesgo

Estas variables son las más predictivas (junto con FP_weighted_mean), ya que cuantifican la masa de riesgo dentro de la mezcla. Los rangos utilizados son muy bajo (menor a 23°C), bajo (entre 23 y 60°C), y alto (mayor a 60°C).

Variable	Significado
prop_fp_muy_bajo	Proporción de masa total de ingredientes cuyo flash point es inferior a 23°C (mayor riesgo).
prop_fp_bajo	Proporción de masa total de ingredientes con flash point entre 23°C y 60°C (riesgo moderado).
prop_fp_alto	Proporción de masa total de ingredientes con flash point mayor a 60°C (seguros).
count_fp_muy_bajo	Conteo de ingredientes con flash point inferior a 23°C.
count_fp_bajo	Conteo de ingredientes con flash point entre 23°C y 60°C.
count_fp_alto	Conteo de ingredientes con flash point mayor a 60°C.

E. Indicadores de calidad de datos e interacciones

Estas variables capturan información sobre la calidad de los datos y combinan conceptos clave.

Variable	Significado
num_ingredientes_censurados	Número de ingredientes en la fórmula que requirieron imputación.
prop_ingredientes_censurados	Proporción de ingredientes censurados sobre el total.
FP_min_x_max_conc	Flash point mínimo multiplicado por la concentración máxima. (Interacción).
num_ing_x_FP_std	Número de ingredientes multiplicado por la desviación estándar del flash point. (Interacción: complejidad versus variabilidad).

El dataset contiene 28.656 filas y 24 columnas luego de aplicar feature engineering y verificar que no existieran filas duplicadas.

Para facilitar el análisis exploratorio y la clasificación de las fórmulas de acuerdo a criterios químicos, este dataset fue dividido en cuatro grupos denominados como compuestos con comportamiento gaseoso (solventes azufrados), soluciones etanólicas, soluciones acuosas, y fórmulas que contienen otros solventes, los cuales corresponden a aldehídos, ésteres y otros compuestos orgánicos.

A cada grupo se le agregó una columna que representa al porcentaje de agua, etanol, solvente, o compuesto gaseoso presente en cada fórmula.

4.2. Metodología

Este Capstone Project abordó la implementación de modelos de machine learning para clasificar 28.656 fórmulas químicas en las categorías muy inflamable, inflamable y no inflamable.

Los datos entregados por Cramer fueron analizados para determinar el tipo de preprocesamiento adecuado para los datasets de fórmulas químicas y de ingredientes.

La variable objetivo (inflamabilidad) fue derivada a partir de los rangos de flash point establecidos por el Decreto 57 del Ministerio de Salud de Chile, categorizando a las fórmulas con flash points inferiores a 23°C como inflamables, fórmulas con flash points con valores iguales o mayores a 23°C e iguales o inferiores a 60°C como inflamables, y todas las fórmulas con flash point superiores a 60°C fueron etiquetadas como no inflamables.

Al revisar el dataset de ingredientes se observó la existencia de valores de tipo NA en la columna Flash Point MP, los cuales fueron imputados con el valor 150°C. Estos valores NA corresponden a compuestos no inflamables.

En el caso del archivo de fórmulas, solamente fue necesario eliminar filas duplicadas antes de proceder a realizar feature engineering para explicar de mejor forma las propiedades fisicoquímicas de las fórmulas, debido a que no se encontraron datos faltantes. En este punto se combinaron ambos datasets para crear 22 variables nuevas, las cuales fueron explicadas en detalle en la sección de datos de este informe.

El siguiente paso fue dividir el dataset completo de fórmulas en cuatro grupos correspondientes a fórmulas que contienen compuestos que presentan comportamiento de tipo gaseoso (azufrados), fórmulas correspondientes a soluciones etanólicas, fórmulas acuosas, y fórmulas que contienen otros tipos de solventes orgánicos, tales como aldehídos y ésteres. En este punto se agregó una columna que representa la proporción de solvente presente en cada fórmula, siendo algunos de ellos de tipo acuoso y los tres grupos restantes contienen solventes de tipo orgánico.

Se procedió a realizar el análisis exploratorio de los cuatro datasets específicos, observando patrones tales como la distribución de los valores de flash point de acuerdo a las tres categorías de inflamabilidad, la distribución de las proporciones de solvente, la distribución de la entropía de Shannon, la cantidad de componentes presentes en las fórmulas, y las correlaciones de las variables por grupo de fórmulas.

Luego de corroborar los patrones presentes en los cuatro datasets se procedió a aplicar modelos de clasificación multiclase (regresión logística multiclase, Random Forest, LightGBM, y CatBoost) calibrados mediante predicción conformal con *margin nonconformity score*, el cual es un tipo de medida de no conformidad o de determinación de lo inusual que es una predicción realizada por un modelo calibrado respecto a las predicciones realizadas anteriormente.

Este puntaje permite determinar la probabilidad de que el clasificador prefiera la clase competidora más probable sobre la verdadera clase de inflamabilidad de una fórmula.

Margin nonconformity score es definido como:

$$s_{margin}(x, y^*) = max_{j \neq y^*} f_j(x) - f_{y^*(x)}$$

donde y^* es la clase verdadera y $f_{y^*(x)}$ es la probabilidad predicha para la clase y^*.

Si $f_{y^*(x)}$ es mucho mayor que las probabilidades de las otras clases, el margen es negativo. Así el ejemplo corresponde a la clase verdadera, por lo tanto, es un resultado conforme. En el caso que el margen sea cercano a cero o positivo, quiere decir que el clasificador no está seguro de la clase o la clase resulta ser incorrecta, por lo tanto, el ejemplo no es conforme (Manokhin, 2025). Este puntaje fue elegido para calibrar los modelos de clasificación porque permite actuar de forma decisiva si la clase resulta ser inflamable o muy inflamable.

La utilización de predicción conformal para calibrar los modelos de machine learning implicó dividir los cuatro datasets en conjuntos de entrenamiento, calibración y prueba. Se consideró un 60% de los datos para entrenamiento, un 20% para calibración y un 20% para prueba para cada dataset.

Los datos solamente fueron estandarizados para el caso de la regresión logística multiclase, utilizando StandardScaler. Este procedimiento no es requerido para los modelos Random Forest, LightGBM, y CatBoost.

Se realizó cross-validation con stratified K-fold, el cual conserva los porcentajes de las muestras para las clases no inflamable, inflamable, y muy inflamable.

Las clases de la variable objetivo fueron convertidas en representaciones numéricas mediante LabelEncoder. Este paso no es requerido para implementar el algoritmo CatBoost, ya que procesa variables categóricas de forma nativa.

Se decidió comparar el rendimiento de dos algoritmos del tipo *gradient boosting* (LightGBM y CatBoost) para comprobar cuál de ellos clasificaba mejor las fórmulas del dataset de la empresa Cramer.

Light Gradient-Boosting Machine (LightGBM) fue desarrollado por Microsoft, mientras que CatBoost fue desarrollado por Yandex. Ambos algoritmos se basan en clasificadores débiles del tipo árbol de decisión, pero difieren en su funcionamiento dado que ocurre un crecimiento a nivel de hoja en el caso de LightGBM, con lo cual puede crear árboles más complejos y profundos que XGBoost. En cambio, el algoritmo CatBoost utiliza árboles de decisión simétricos y realiza refuerzo ordenado para evitar el riesgo de sobreajuste que puede ocurrir al usar algoritmos como LightGBM y XGBoost. Además, LightGBM es más veloz computacionalmente que XGBoost.

Las métricas consideradas para evaluar el rendimiento de los modelos regresión logística multiclase, Random Forest, LightGBM, y CatBoost fueron accuracy, precision, recall, y F1-score, graficando los resultados mediante matrices de confusión.

A continuación se analizaron los errores detectados en los cuatro modelos de clasificación, enfocándose en los que correspondía a la clasificación de fórmulas muy inflamables como no inflamables, y viceversa.

El paso final del estudio de los modelos de clasificación multiclase fue aplicar el método SHAP para interpretar la importancia de las características del modelo que obtuvo el mejor rendimiento general para los cuatro grupos de fórmulas.

5. Resultados

Análisis exploratorio

Los cuatro conjuntos de fórmulas presentan un desbalance de clases de acuerdo a lo observado en la Figura N°1. Solamente en el caso de las soluciones etanólicas predomina la clase muy inflamable, mientras que la clase no inflamable es la más común en los otros tres conjuntos de datos. Esto influirá en la capacidad de los modelos de discriminar adecuadamente entre las tres clases de inflamabilidad.

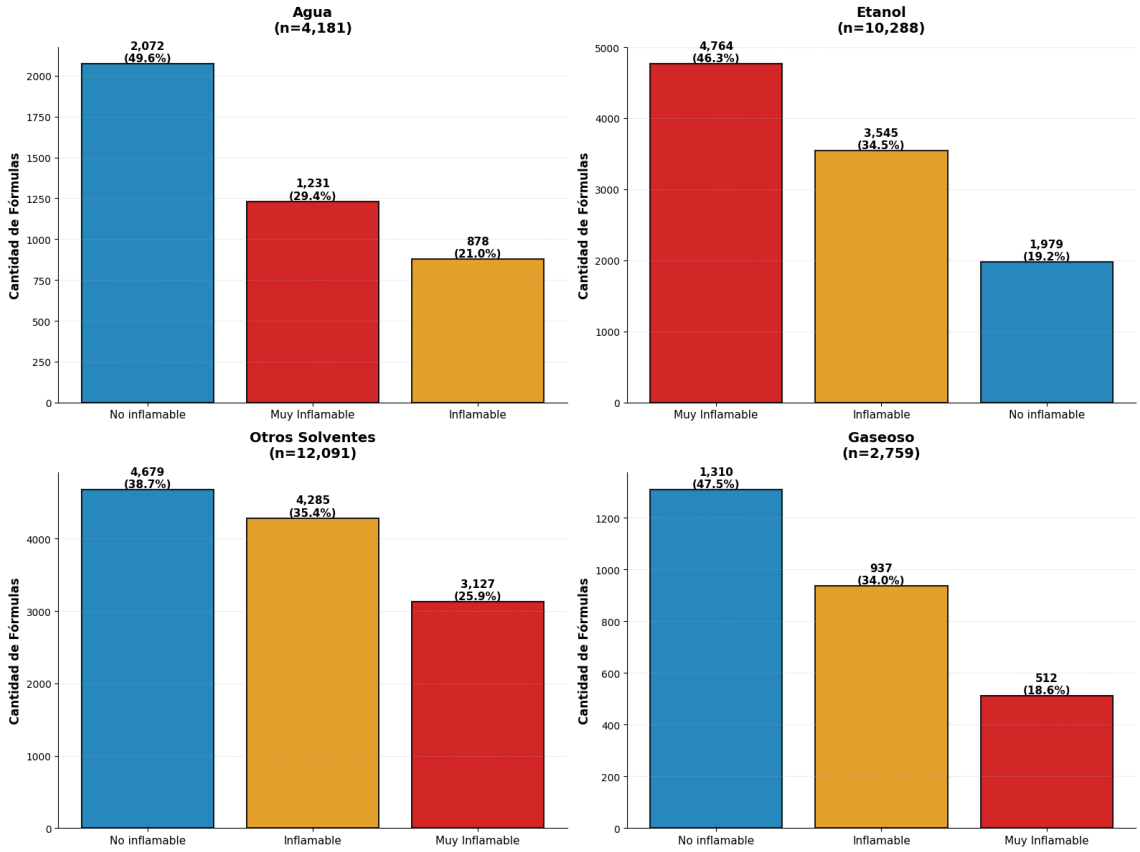

Figura N°1: Distribución de las clases no inflamable, muy inflamable, e inflamable por grupo de fórmulas.

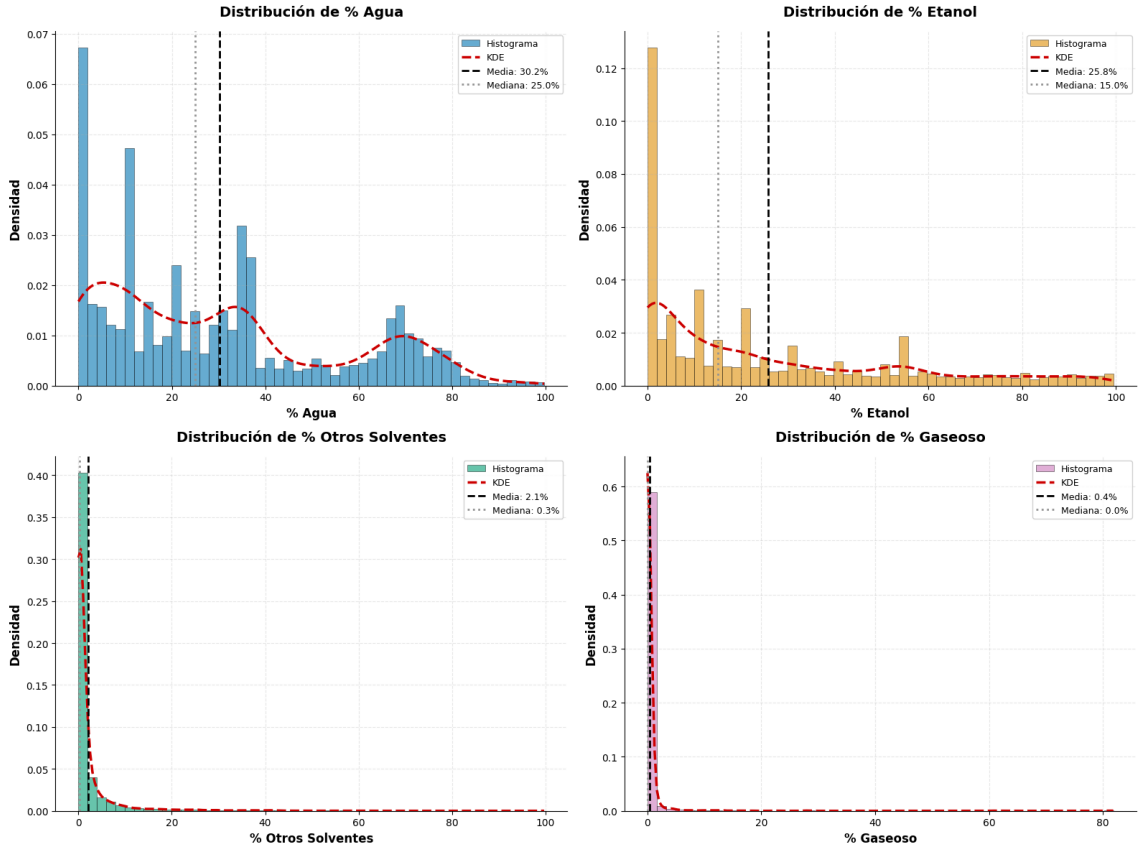

Figura N°2: Distribución de los porcentajes de solvente por grupo de fórmula.

En la Figura N°2 se comprueba que los porcentajes de solvente se concentran en valores bajos en el caso de las fórmulas que contienen otros solventes y en las que presentan compuestos con comportamiento gaseoso. Este fenómeno se observa en los otros grupos, pero existe una mayor dispersión en el caso de las soluciones acuosas y soluciones etanólicas. Esto concuerda con los gráficos de la Figura N°3, el cual muestra que las fórmulas con compuestos de tipo gaseoso presentan el menor rango intercuartílico en las tres clases estudiadas, mientras que ocurre lo contrario con las soluciones acuosas y la clase muy inflamable de las soluciones etanólicas.

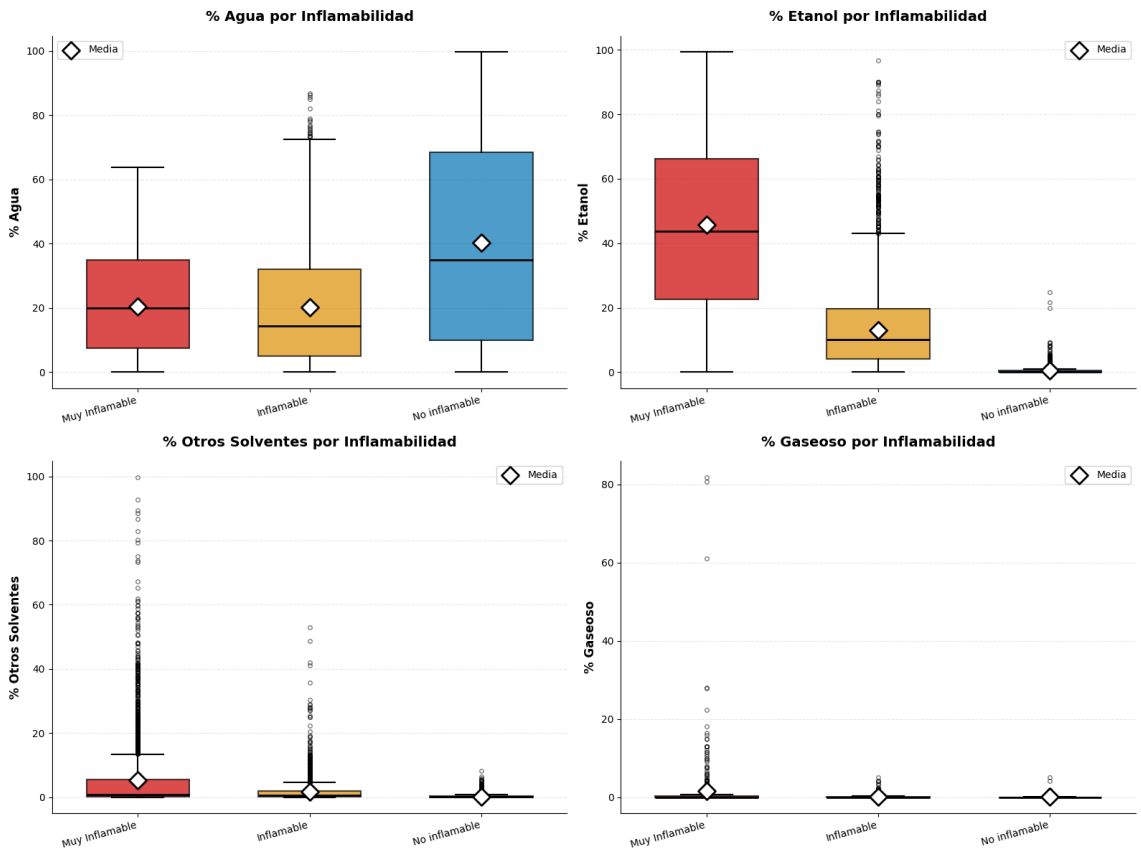

Figura N°3: Distribución de los porcentajes de solvente por grupo de fórmula, incluyendo sus valores atípicos.

Los otros grupos que presentan rangos intercuartílicos pequeños son las clases no inflamable de las fórmulas etanólicas y las que contienen otros solventes orgánicos. La clase muy inflamable de las fórmulas compuestas por otros solventes presenta una gran cantidad de valores atípicos.

Hasta este punto del análisis exploratorio se confirmaría la importancia de separar las fórmulas de acuerdo a su composición química.

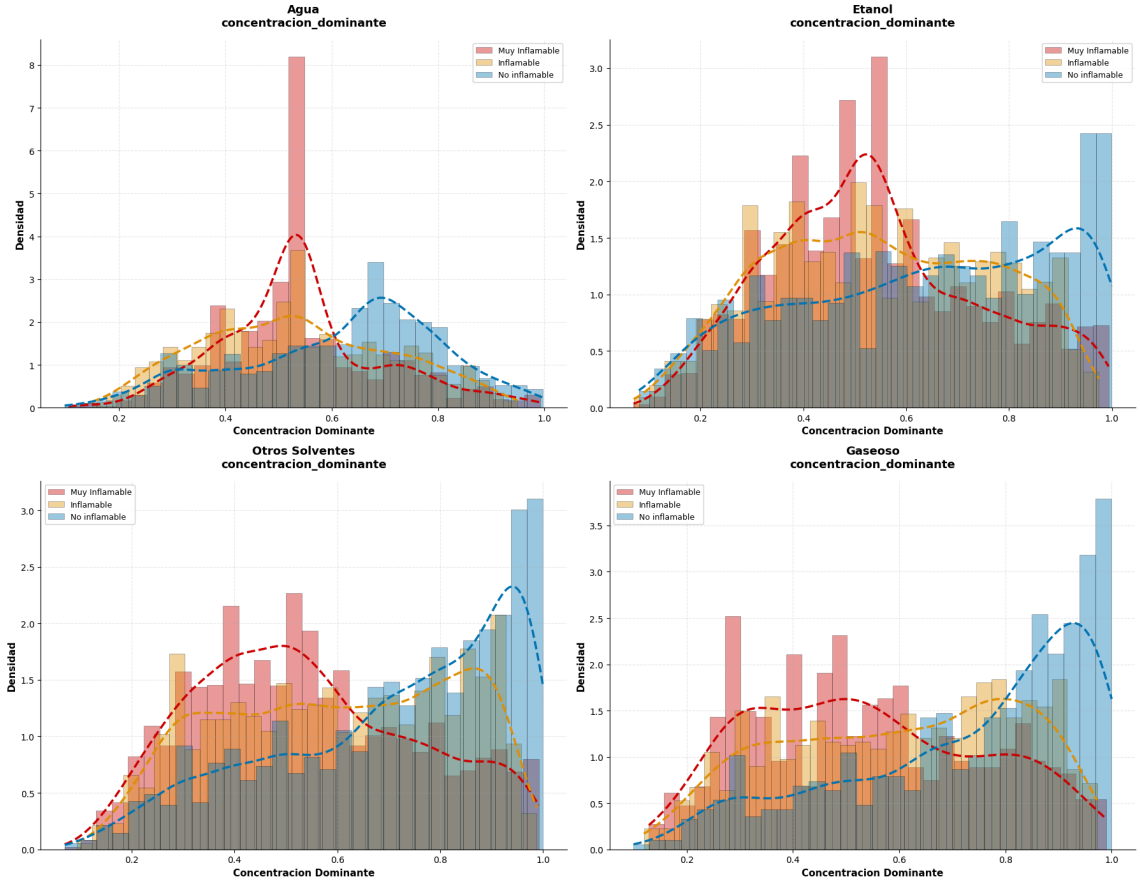

Figura N°4: Distribución de los valores de concentración dominante de las fórmulas analizadas.

Se observan asimetrías negativas en el caso de las clases no inflamables de las mezclas que contienen otros solventes y compuestos con comportamiento de tipo gaseoso.

La clase muy inflamable del grupo de soluciones acuosas presenta un pico aproximadamente a una concentración igual a 0,5. Un patrón similar es detectado en la clase muy inflamable de las soluciones etanólicas.

En general, todas las clases de inflamabilidad presentan distribuciones características en los gráficos correspondientes a los cuatro grupos estudiados.

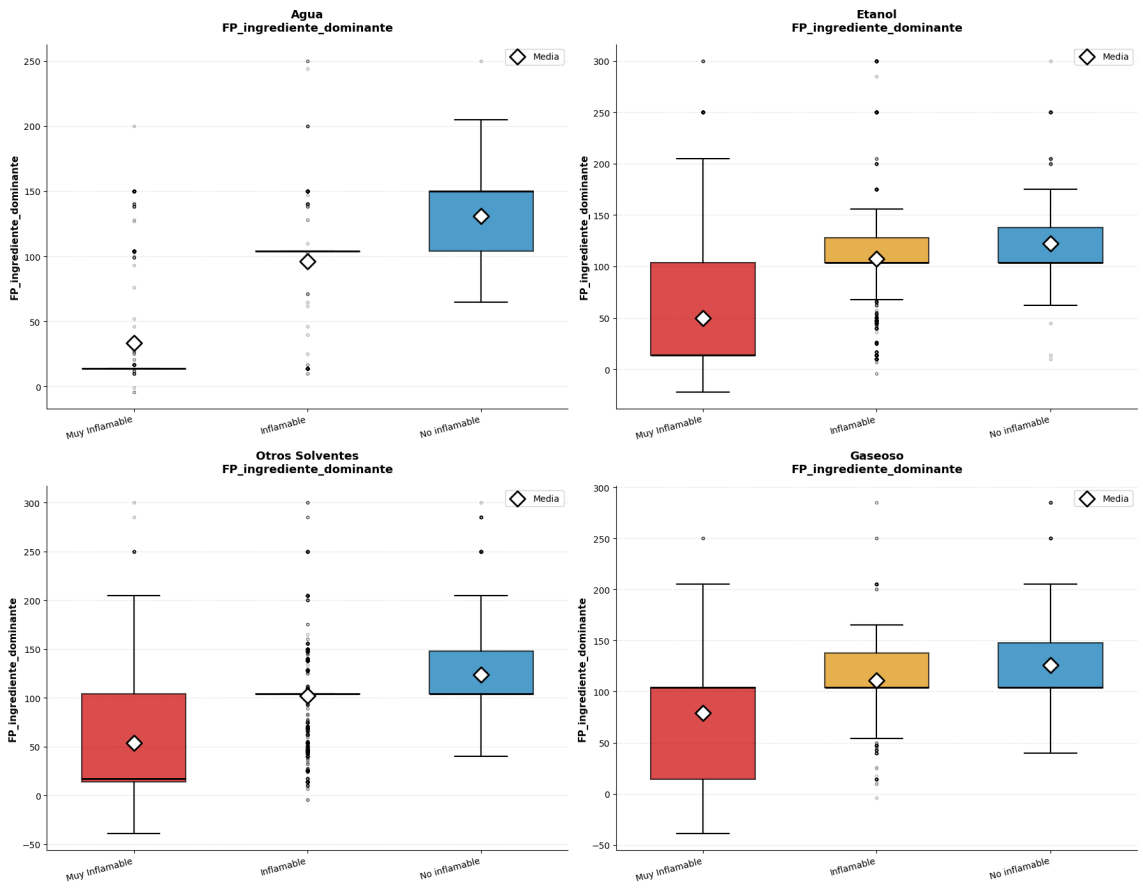

Figura N°5: Distribución del flashpoint del ingrediente dominante de las fórmulas, incluyendo sus valores atípicos.

Los menores rangos intercuartílicos de las distribuciones de los valores de flash point del ingrediente dominante de una fórmula corresponden a las clases muy inflamable e inflamable de las soluciones acuosas, y las fórmulas inflamables que contienen otros solventes orgánicos. El caso contrario se observa en las clases muy inflamable de las soluciones etanólicas, compuestas por otros solventes orgánicos, y las fórmulas que contienen compuestos con comportamiento gaseoso. Los menores valores mínimos de flash point del ingrediente dominante de una fórmula se encuentran en las clases muy

inflamables de las fórmulas compuestas por ingredientes con comportamiento gaseoso y que contienen solventes tales como aldehídos y ésteres, siendo cercano a -50 °C.

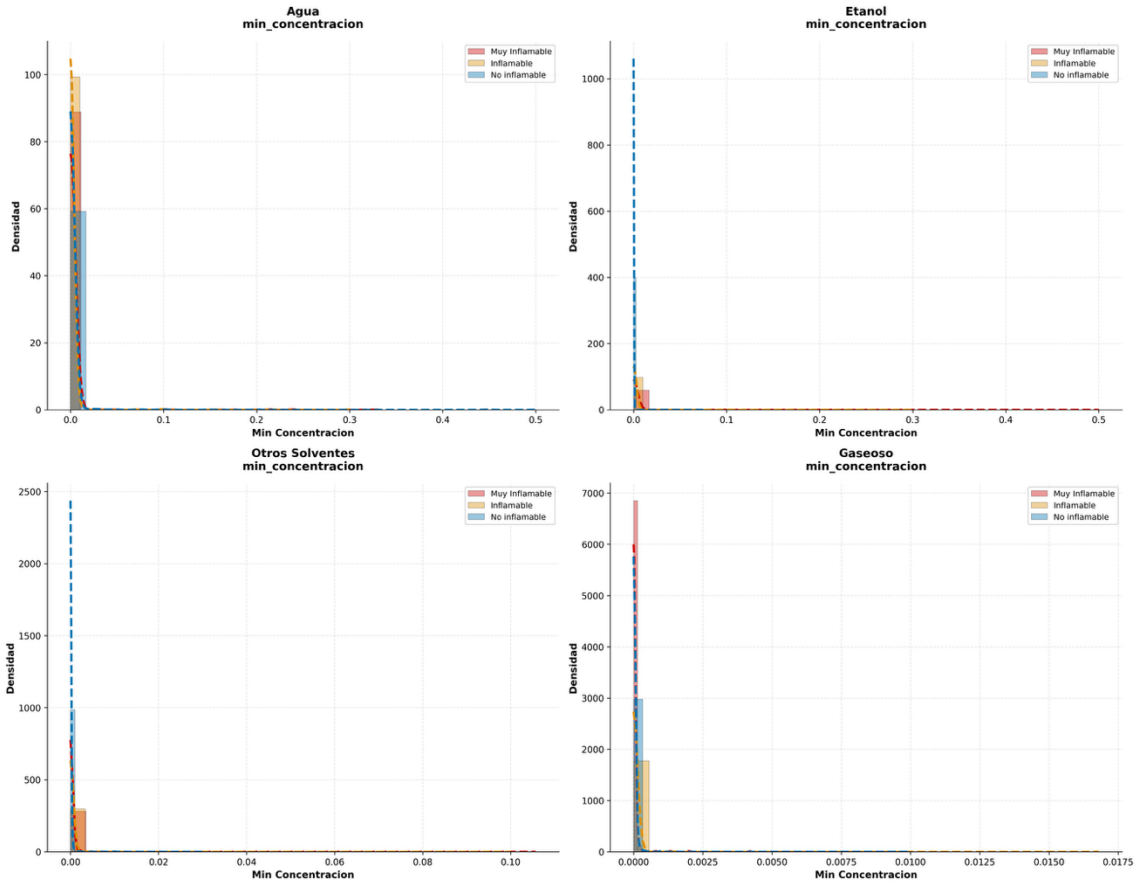

Figura N°6: Distribución de las concentraciones mínimas de las fórmulas analizadas.

Las menores concentraciones de agua son más frecuentes en el caso de los compuestos acuosos inflamables y muy inflamables. El grupo que concentra una mayor cantidad de fórmulas en este gráfico son las fórmulas que contienen compuestos con comportamiento de tipo gaseoso, seguido por las fórmulas que contienen otros solventes orgánicos con bajo flash point.

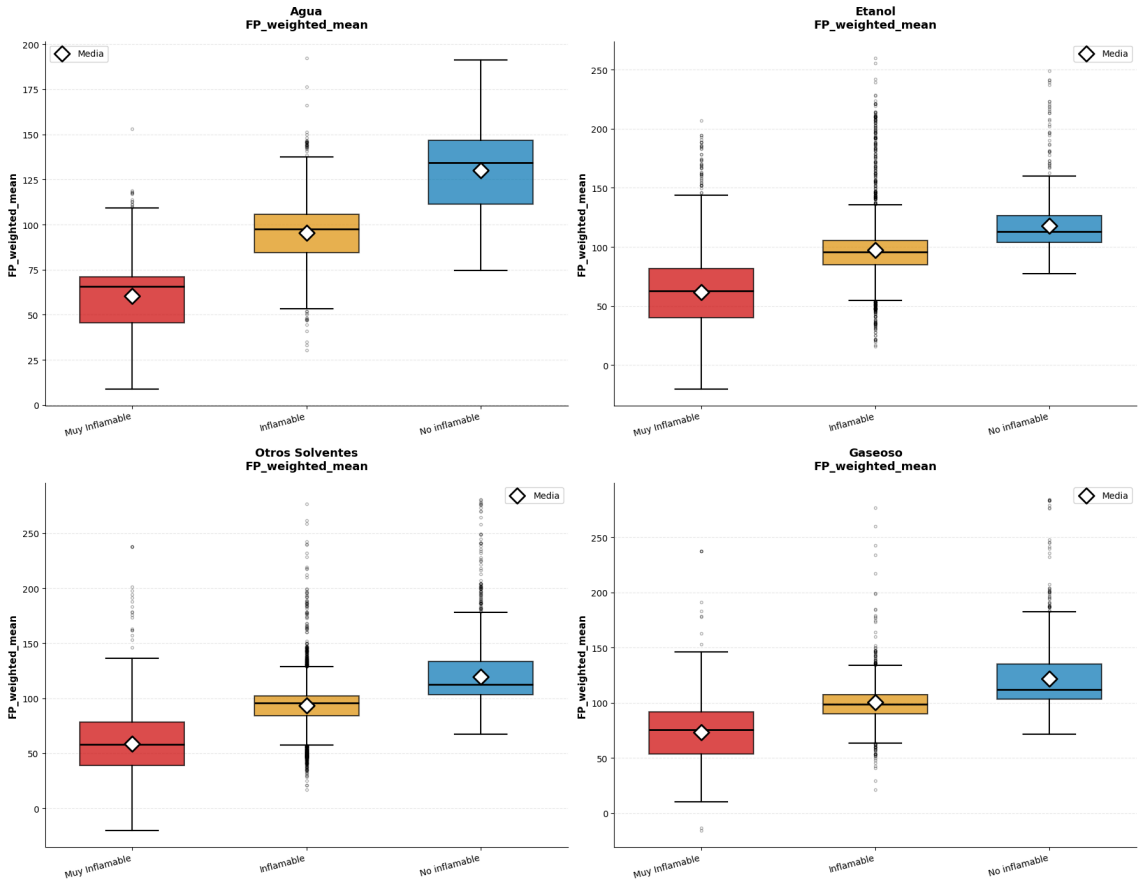

Figura N°7: Distribución del promedio ponderado del flash point de las fórmulas analizadas, incluyendo valores atípicos.

En todos los grupos se verifica que el promedio ponderado del flash point de distribuye a valores más altos en el caso de las fórmulas no inflamables, y los menores valores corresponden a fórmulas muy inflamables.

Existe una gran cantidad de valores atípicos en los grupos inflamables de las soluciones etanólicas y de las fórmulas que contienen otro tipo de solventes orgánicos.

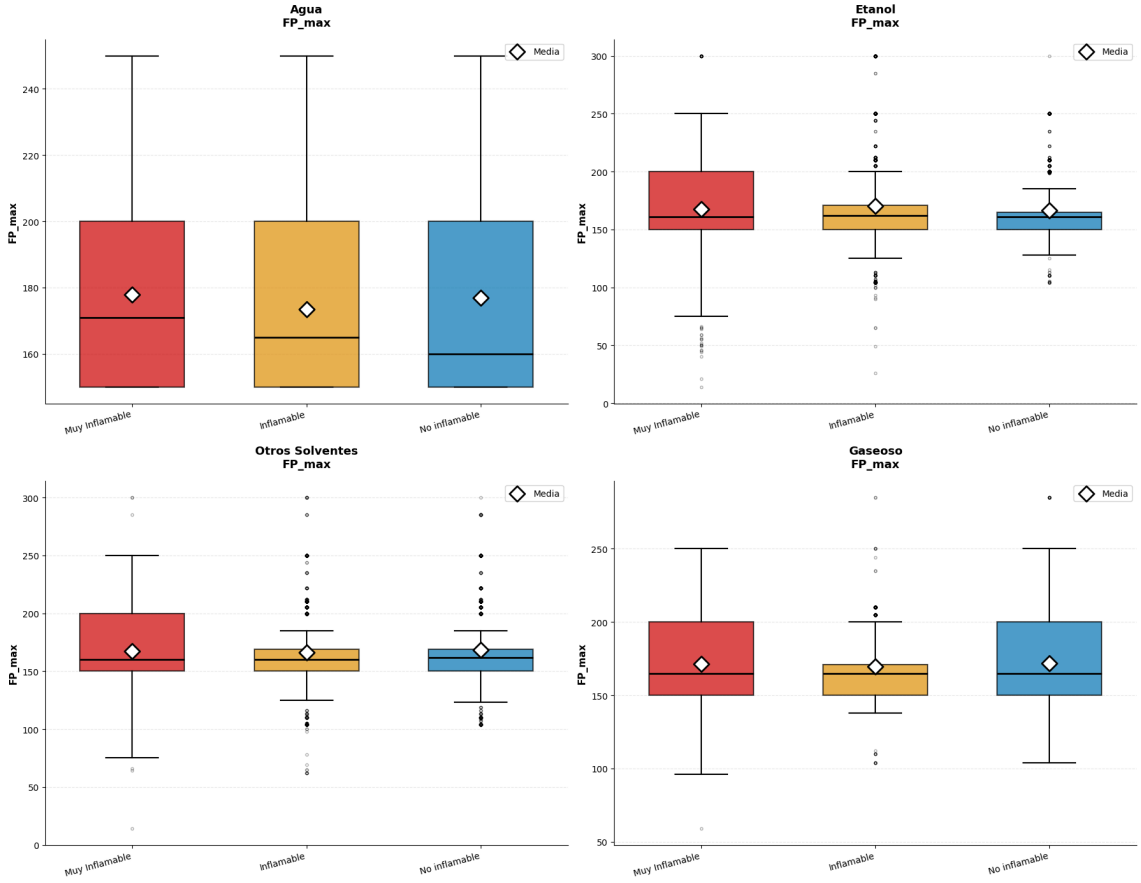

Figura N°8: Distribución del flash point máximo de las fórmulas analizadas, incluyendo valores atípicos.

Las fórmulas acuosas presentan consistentemente flash points máximos iguales o superiores a 150°C. El caso contrario se observa para las fórmulas muy inflamables de los otros grupos, siendo los menores valores detectados en el conjunto de fórmulas que contienen otros solventes con bajo flash point.

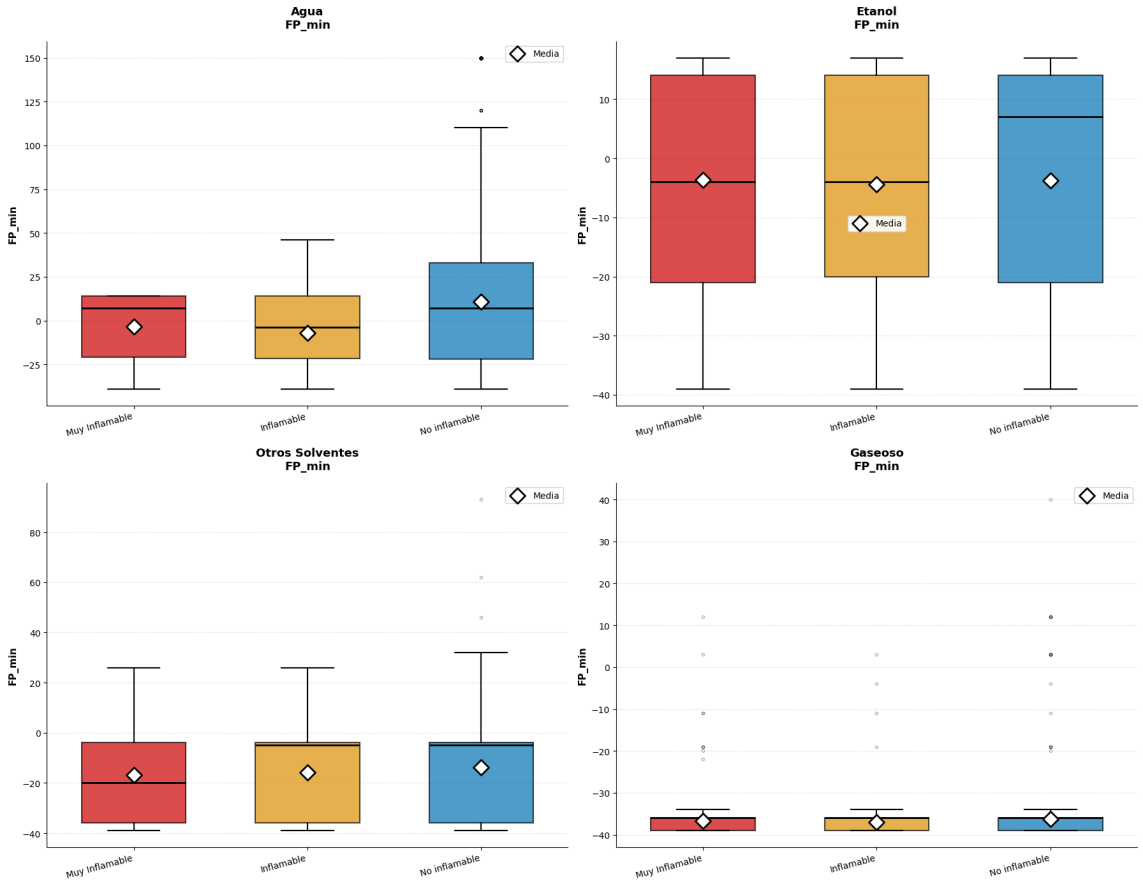

Figura N°9: Distribución del flash point mínimo de las fórmulas analizadas, incluyendo valores atípicos.

Todos los grupos presentan flash points mínimos con valores negativos de temperatura, siendo más bajos en el caso de los tres grupos que contienen solventes orgánicos.

Las fórmulas con compuestos con comportamiento de tipo gaseoso tienen una distribución de sus flash points mínimos con rangos intercuartílicos muy estrechos. Todos los valores cercanos o superiores a -20°C son considerados como valores atípicos para este grupo.

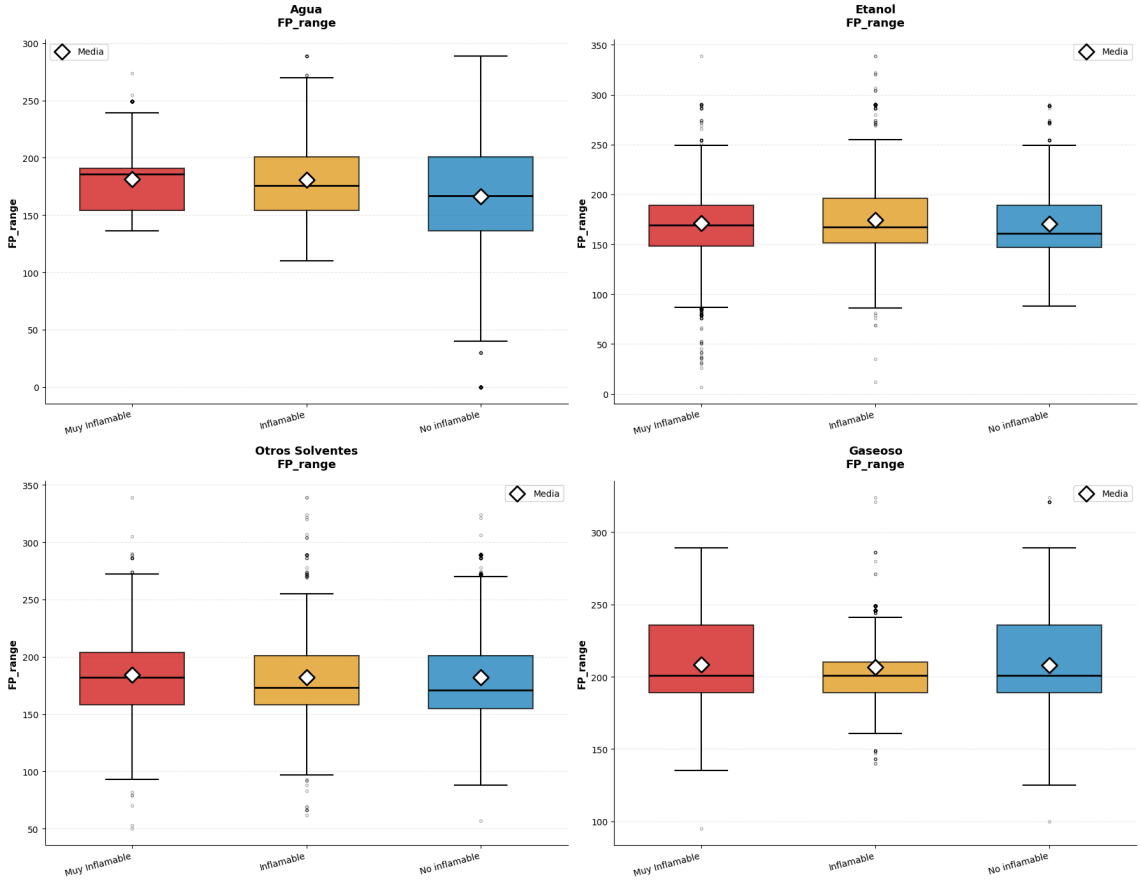

Figura N°10: Distribución del rango de flash point de las fórmulas analizadas, incluyendo valores atípicos.

Las fórmulas no inflamables acuosas muestran un mayor rango intercuartílico en comparación a las fórmulas inflamables y muy inflamables. En cambio, los rangos intercuartílicos de las fórmulas que contienen otros solventes son similares entre sí.

El menor rango intercuartílico se observa en el grupo inflamable de las fórmulas que contienen ingredientes con comportamiento gaseoso.

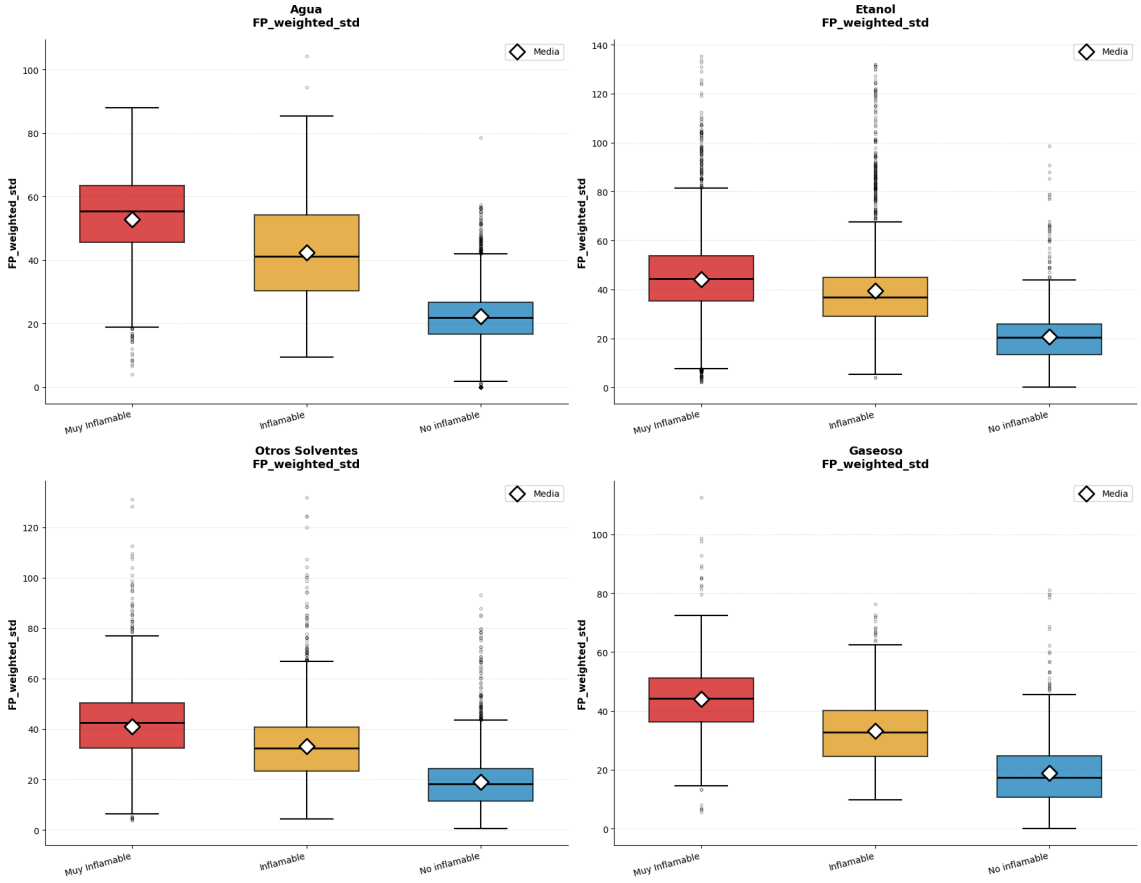

Figura N°11: Distribución del promedio ponderado de la desviación estándar del flash point de las fórmulas analizadas, incluyendo valores atípicos.

En tres grupos de fórmulas se confirma que el rango intercuartílico de la distribución del promedio ponderado de la desviación estándar es menor en las fórmulas no inflamables. La excepción es el grupo de fórmulas con compuestos con comportamiento de tipo gaseoso, dado que los rangos intercuartílicos de las tres clases de inflamabilidad son muy parecidos entre sí. Se observa una gran cantidad de valores atípicos en la clase inflamable de las fórmulas etanólicas.

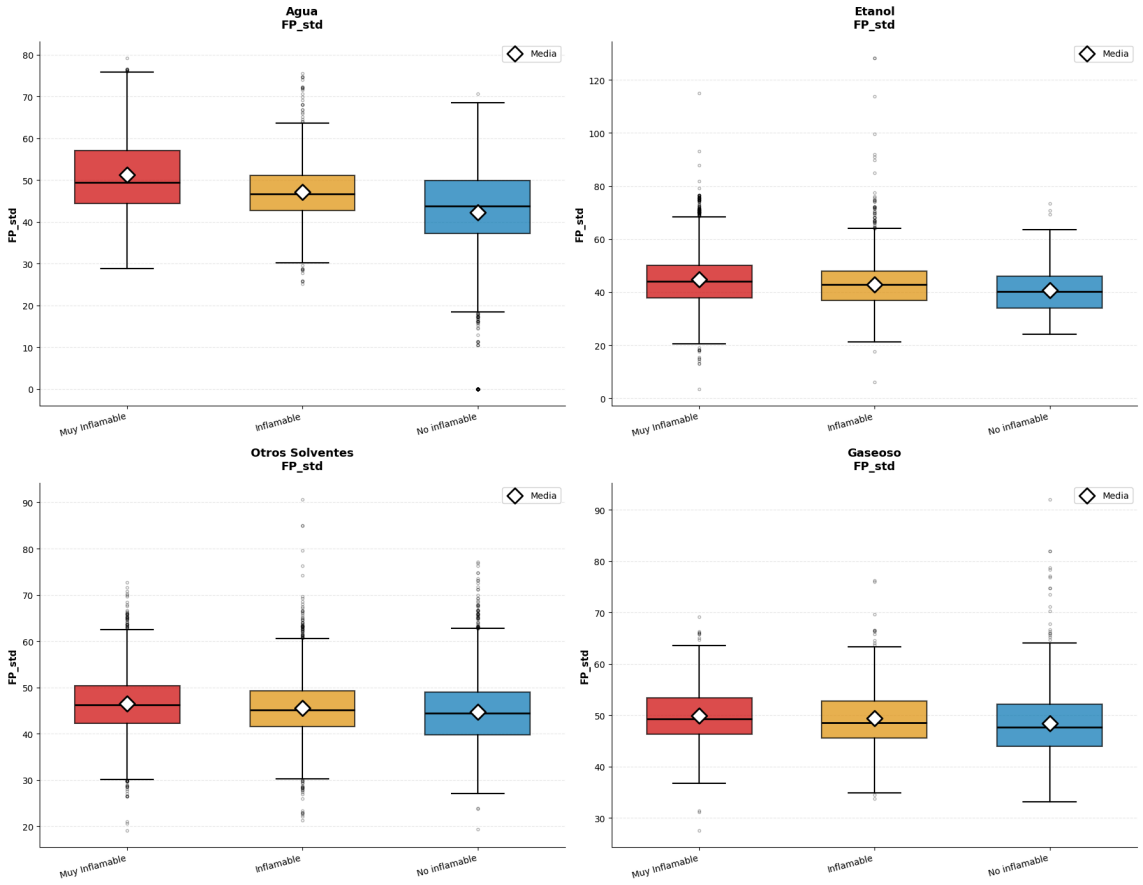

Figura N°12: Distribución de la desviación estándar del flash point de las fórmulas analizadas, incluyendo valores atípicos.

Los rangos intercuartílicos de las clases de inflamabilidad de las fórmulas que contienen solventes orgánicos son similares entre sí, lo cual no ocurre con la clase inflamable de las fórmulas acuosas al ser comparado con las fórmulas muy inflamables y no inflamables.

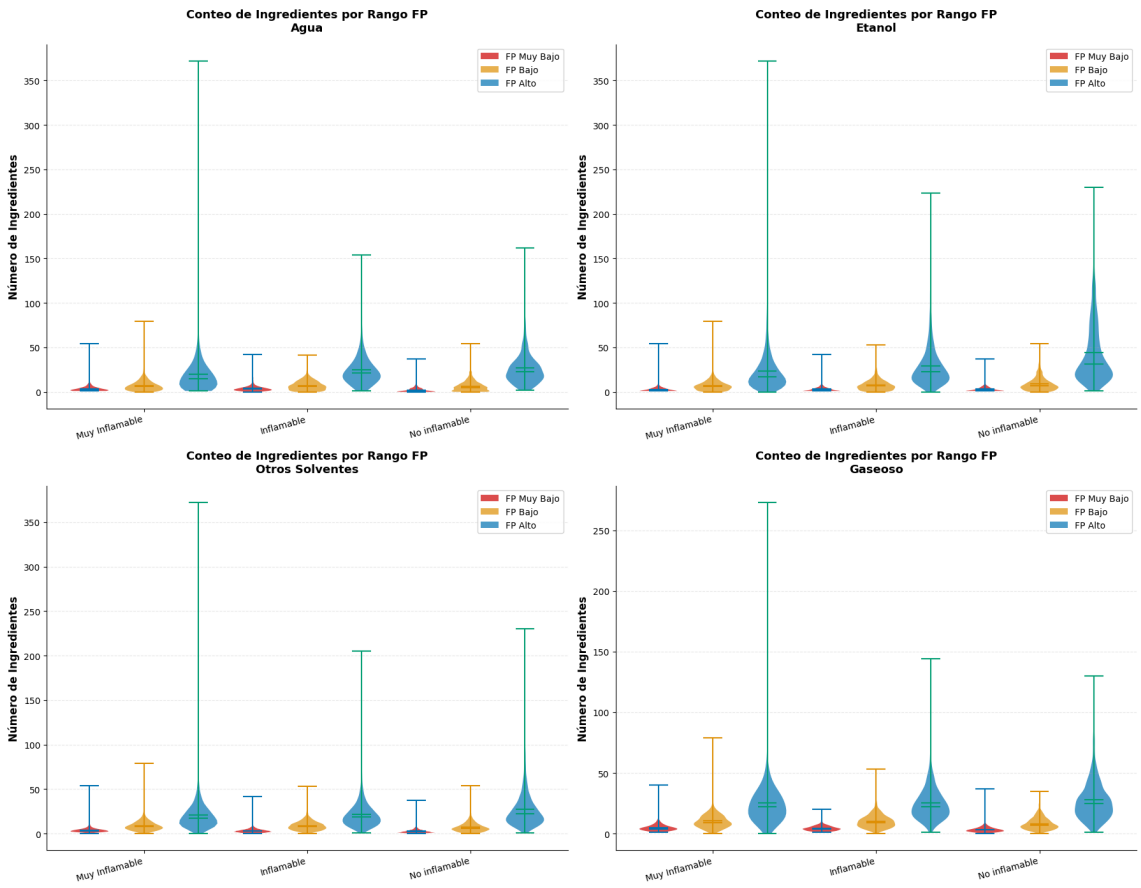

Figura N°13: Distribución de la cantidad de ingredientes contenidos en las fórmulas analizadas.

En todos los grupos de fórmulas se comprueba que existe una mayor cantidad de ingredientes de flash point alto, seguido de ingredientes de flash point bajo, y la menor cantidad corresponde a ingredientes de flash point muy bajo.

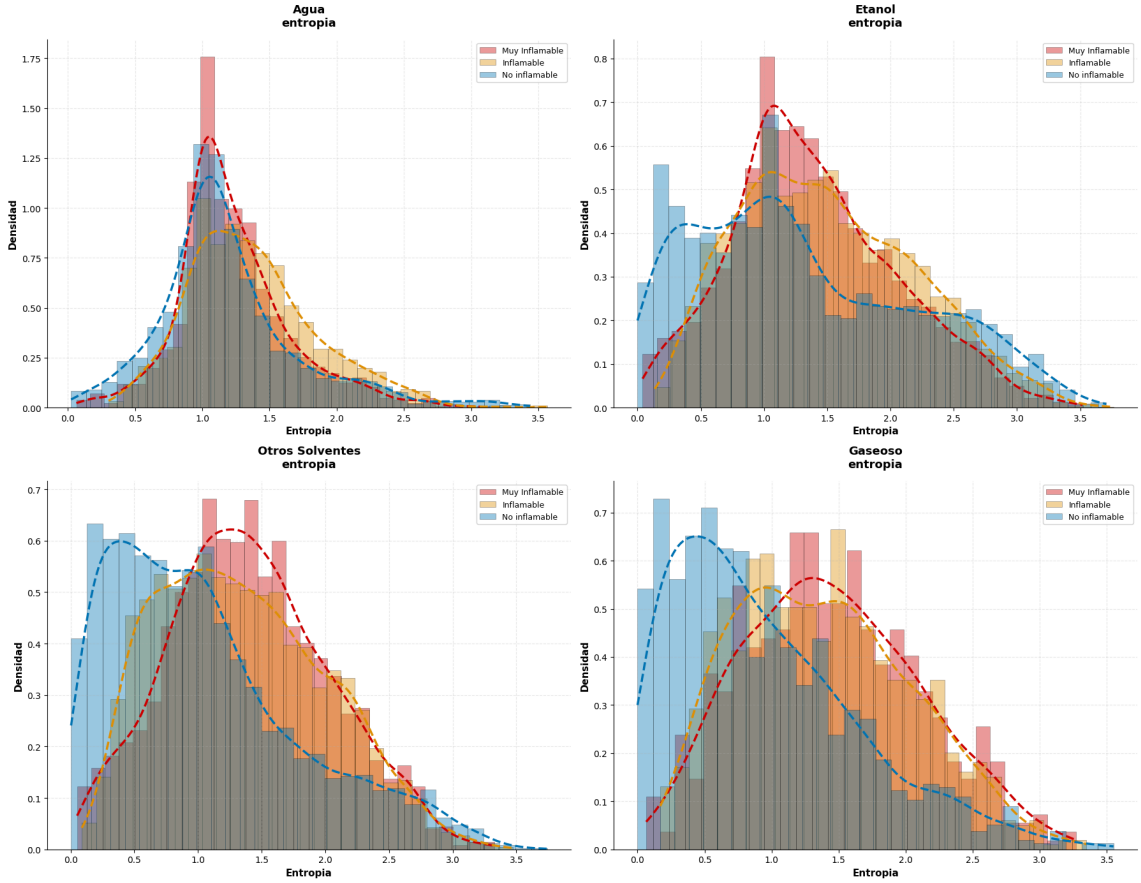

Figura N°14: Distribución de la entropía de Shannon de las fórmulas analizadas.

La entropía de Shannon mide la uniformidad de la distribución de las concentraciones de una fórmula química.

Las distribuciones de las fórmulas muy inflamables y no inflamables presentan sus picos en aproximadamente el mismo valor en el caso de las fórmulas acuosas.

En los cuatro grupos se comprueba que las distribuciones de la entropía de Shannon son características para las clases de inflamabilidad correspondientes. Esto refuerza la necesidad de evaluar cada grupo por separado.

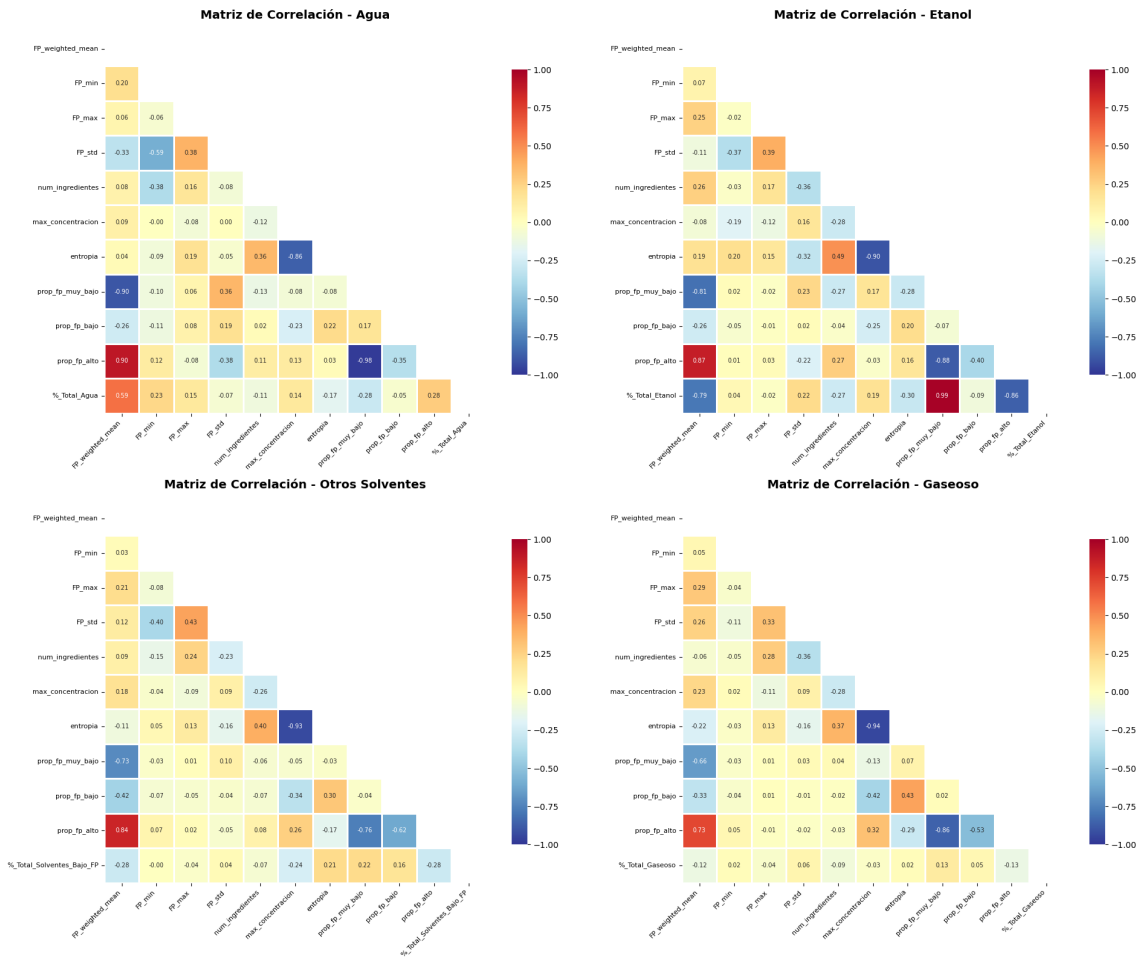

Figura N°15: Correlación de las variables correspondientes a fórmulas acuosas, etanólicas, que contienen otros tipos de solventes, y que contienen ingredientes con comportamiento gaseoso.

Se comprueba que los patrones de correlaciones entre variables son diferentes para cada tipo de fórmula química. Este antecedente concuerda con lo observado en las otras variables consideradas en el análisis exploratorio. Por lo tanto, se puede adelantar que los cuatro grupos fueron modelados por separado en la etapa de aplicación de modelos de machine learning.

Modelos de machine learning de clasificación multiclase

Regresión logística multiclase con predicción conformal al 90% de confianza

Los modelos de regresión logística multiclase fueron calibrados mediante predicción conformal, considerando un 90% de confianza. En la tabla N°1 se comparan las coberturas empíricas de los cuatro grupos de fórmulas, verificando que los grupos correspondientes a fórmulas que contienen ingredientes con comportamiento de tipo gaseoso y solventes con bajo flash point se obtuvo una cobertura inferior a 0,90. Estos resultados confirman la necesidad de probar otros modelos de clasificación multiclase para verificar si esta cobertura puede ser mejorada.

Tipo de fórmula	Cobertura empírica
Acuosa	0,904
Etanólica	0,913
Comportamiento tipo gaseoso	0,893
Otros solventes	0,893

Tabla N°1: Cobertura empírica para regresión logística multiclase. El valor esperado para cada grupo es igual a 0,90.

En la Figura N°16 se comparan las matrices de confusión de los cuatro grupos de fórmulas. En el caso de las fórmulas con ingredientes con comportamiento de tipo gaseoso no ocurrieron errores de clasificación muy graves, es decir, no se clasificaron fórmulas muy inflamables como no inflamables, y viceversa. En los otros grupos se etiquetaron como no inflamables dos fórmulas acuosas, tres fórmulas etanólicas, y cinco fórmulas que contienen otros solventes, a pesar que se trataba realmente de fórmulas muy inflamables.

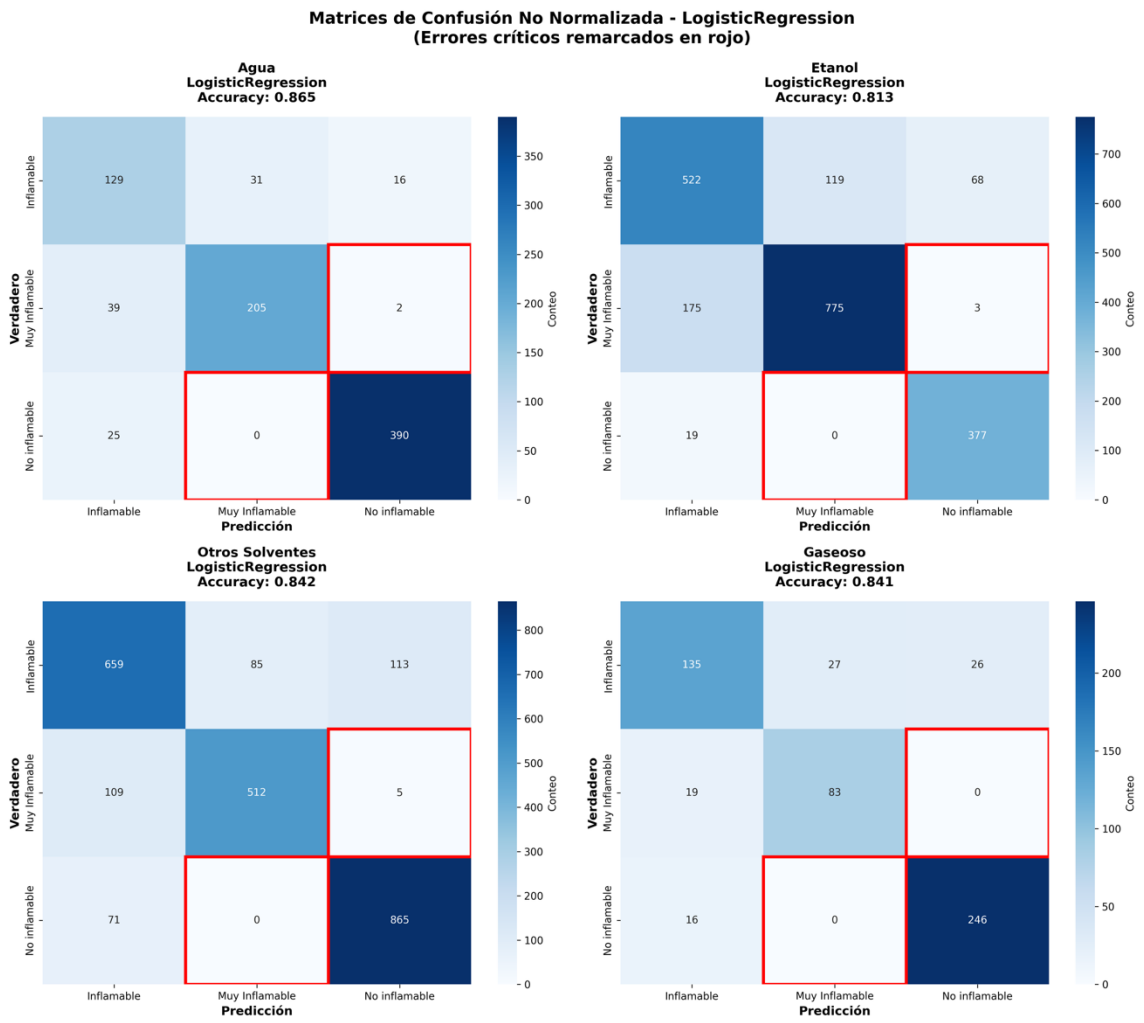

Figura N°16: Matrices de confusión para los grupos de fórmulas clasificados por el modelo regresión logística multiclase según su composición química.

En la Figura N°17 observamos que en general el grupo con mayor F1-score corresponde a las fórmulas que contienen otros solventes, mientras que los valores más bajos de esta métrica se obtuvieron en los grupos de fórmulas etanólicas y con ingredientes con comportamiento de tipo gaseoso. Estas métricas son explicadas en mayor detalle en la Tabla N°2.

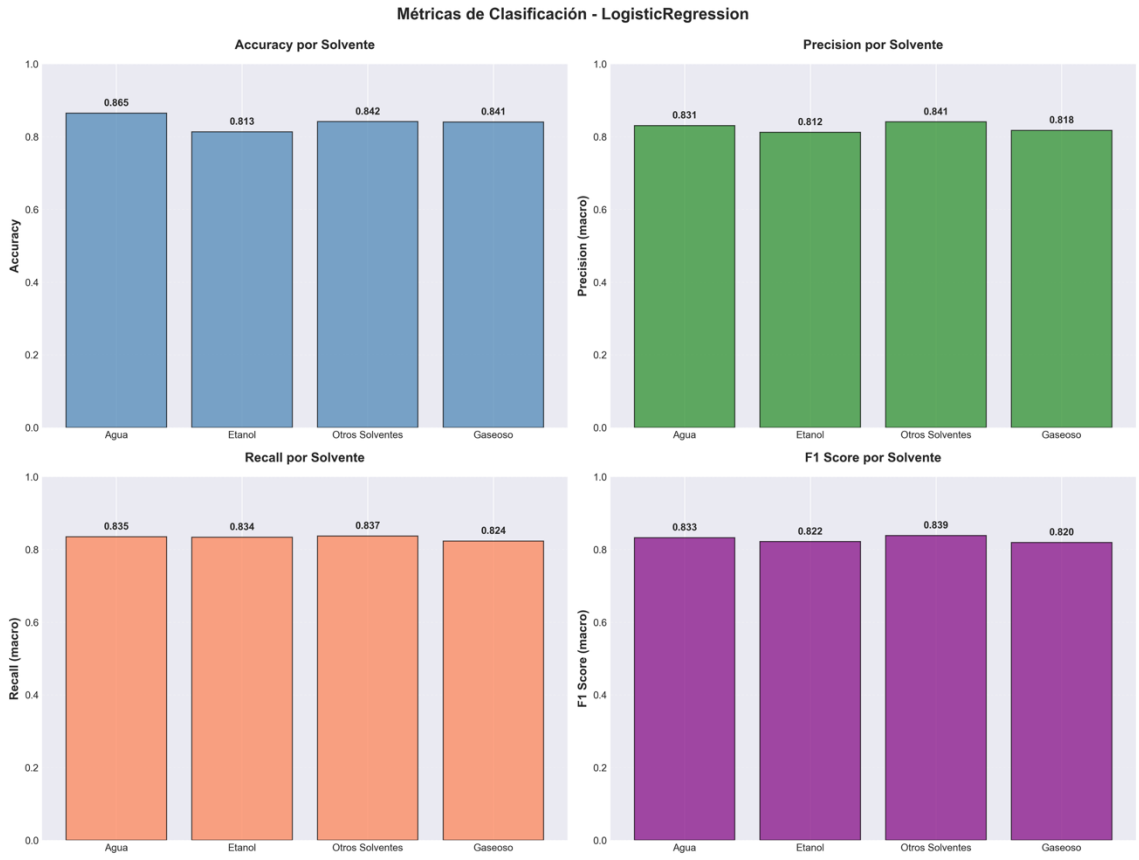

Figura N°17: Métricas de rendimiento del modelo regresión logística multiclase.

Consistentemente podemos observar en la Tabla N°2 que en los cuatro grupos los peores valores de F1-score son obtenidos por la clase inflamable. Lo contrario ocurre en el caso de la clase no inflamable. Es posible que los modelos de regresión logística multiclase tengan mayor dificultad para distinguir la clase inflamable debido a que posee valores intermedios, en comparación a las clases extremas muy inflamable y no inflamable. Además, en general son más comunes las fórmulas no inflamables, salvo en el caso de las fórmulas etanólicas, en las cuales predomina la clase muy inflamable.

Grupo fórmula	Accuracy	Precision	Recall	F1-score
Acuoso muy inflamable	0,833	0,869	0,833	0,851
Acuoso inflamable	0,733	0,668	0,733	0,699
Acuoso no inflamable	0,940	0,956	0,940	0,948
Etanólico muy inflamable	0,813	0,867	0,813	0,839
Etanólico inflamable	0,736	0,729	0,736	0,733
Etanólico no inflamable	0,952	0,842	0,952	0,893
Gaseoso muy inflamable	0,814	0,755	0,814	0,783
Gaseoso inflamable	0,718	0,794	0,718	0,754
Gaseoso no inflamable	0,939	0,904	0,939	0,921
Otros solventes muy inflamable	0,818	0,857	0,818	0,837
Otros solventes inflamable	0,769	0,785	0,769	0,777
Otros solventes no inflamable	0,924	0,880	0,924	0,902

Tabla N°2: Métricas clasificación regresión logística multiclase.

Random Forest con predicción conformal al 90% de confianza

Los modelos Random Forest fueron calibrados mediante predicción conformal, considerando un 90% de confianza. En la Tabla N°3 se comparan las coberturas empíricas de los cuatro grupos de fórmulas, observando que en los cuatro grupos se obtuvo una cobertura empírica igual o superior a 0,90.

Tipo de fórmula	Cobertura empírica
Acuosa	0,913
Etanólica	0,903
Comportamiento tipo gaseoso	0,920
Otros solventes	0,904

Tabla N°3: Cobertura empírica para Random Forest. El valor esperado para cada grupo es igual a 0,90.

En la Figura N°18 se comparan las matrices de confusión de los cuatro grupos de fórmulas. En el caso de las fórmulas etanólicas no ocurrieron errores de clasificación muy graves, es decir, no se clasificaron fórmulas muy inflamables como no inflamables, y viceversa. En el caso de las fórmulas con comportamiento de tipo gaseoso, se etiquetó incorrectamente una fórmula no inflamable como muy inflamable, error que no ocurrió en la regresión logística multiclase. Además, una fórmula acuosa y tres fórmulas que contienen solventes con bajo flash point fueron asignadas como fórmulas no inflamables, siendo realmente fórmulas muy inflamables.

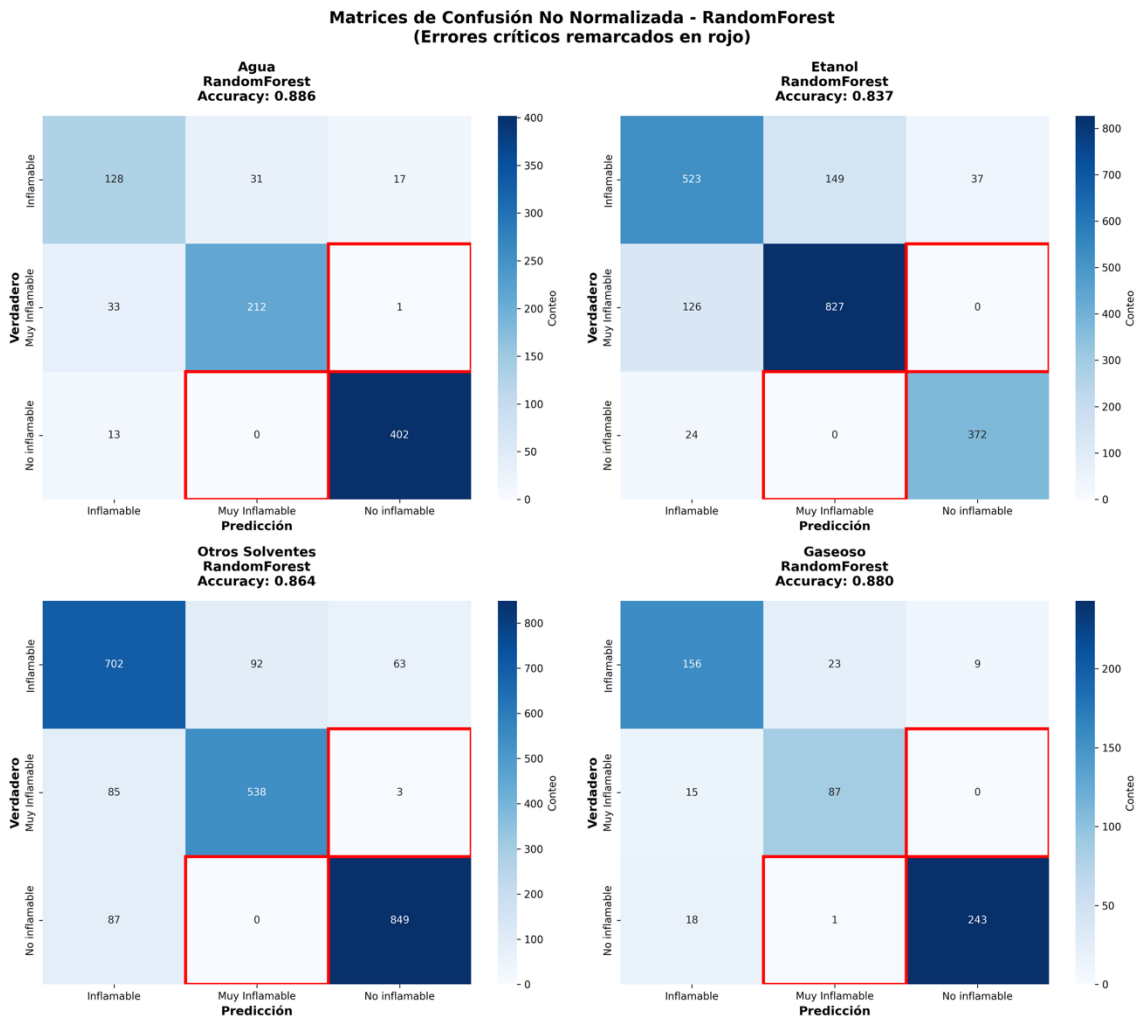

Figura N°18: Matrices de confusión para los grupos de fórmulas clasificados por el modelo Random Forest según su composición química.

Respecto a las métricas de los modelos Random Forest, en la Figura N°19 podemos verificar que los mayores valores de F1-score fueron obtenidos para los grupos de fórmulas con ingredientes con comportamiento de tipo gaseoso y fórmulas que contienen solventes orgánicos con bajo flash point. El rendimiento más bajo fue logrado

por las fórmulas etanólicas. Al comparar los valores con los modelos de regresión logística multiclase, todos los resultados fueron mejores en comparación.

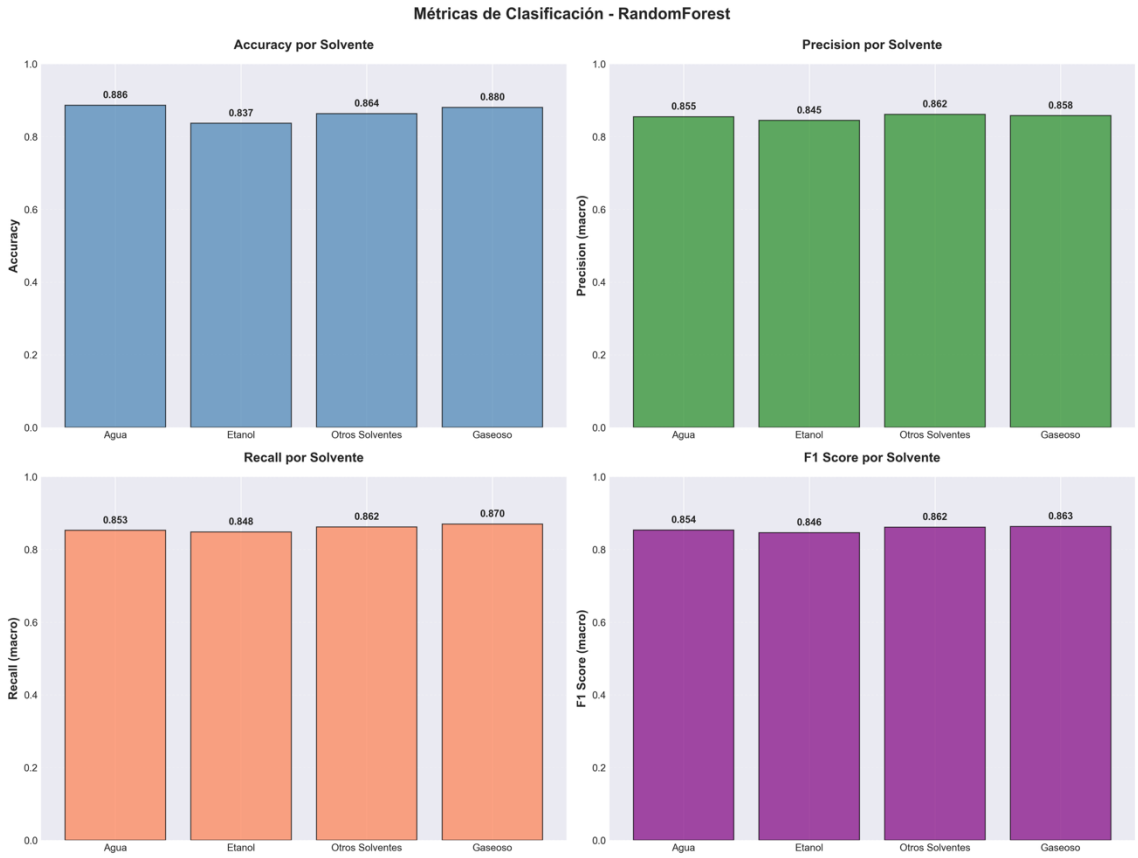

Figura N°19: Métricas de rendimiento del modelo Random Forest.

En la Tabla N°4 se muestra el desglose de las métricas de rendimiento de los modelos Random Forest por clase de inflamabilidad y grupo de fórmula. Al igual que en el caso de los modelos de regresión logística, la clase con menor rendimiento es la inflamable, mientras que la clase no inflamable tiene mejores métricas en los cuatro grupos de fórmulas.

Grupo fórmula	Accuracy	Precision	Recall	F1-score
Acuoso muy inflamable	0,862	0,872	0,862	0,867
Acuoso inflamable	0,727	0,736	0,727	0,731
Acuoso no inflamable	0,969	0,957	0,969	0,963
Etanólico muy inflamable	0,868	0,847	0,868	0,857
Etanólico inflamable	0,738	0,777	0,738	0,757
Etanólico no inflamable	0,939	0,910	0,939	0,924
Gaseoso muy inflamable	0,853	0,784	0,853	0,817
Gaseoso inflamable	0,830	0,825	0,830	0,828
Gaseoso no inflamable	0,927	0,964	0,927	0,946
Otros solventes muy inflamable	0,859	0,854	0,859	0,857
Otros solventes inflamable	0,819	0,803	0,819	0,811
Otros solventes no inflamable	0,907	0,928	0,907	0,917

Tabla N°4: Métricas clasificación Random Forest.

LightGBM con predicción conformal al 90% de confianza

Los modelos LightGBM fueron calibrados mediante predicción conformal, considerando un 90% de confianza. En la tabla N°5 se comparan las coberturas empíricas de los cuatro grupos de fórmulas, comprobando que no se logró el valor esperado para las soluciones etanólicas y las fórmulas con ingredientes que presentan comportamiento de tipo gaseoso.

Tipo de fórmula	Cobertura empírica
Acuosa	0,931
Etanólica	0,893
Comportamiento tipo gaseoso	0,893
Otros solventes	0,902

Tabla N°5: Cobertura empírica para LightGBM. El valor esperado para cada grupo es igual a 0,90.

En la Figura N°20 se comparan las matrices de confusión de los cuatro grupos de fórmulas. En el caso de las fórmulas acuosas y de las fórmulas etanólicas se clasificó respectivamente una de ellas como no inflamable, cuando en realidad se trataba de dos fórmulas muy inflamables. Este mismo tipo de error ocurrió con tres fórmulas que contienen solventes con bajo flash point.

En el caso de las fórmulas con comportamiento gaseoso, se etiquetó erróneamente una fórmula no inflamable como muy inflamable, error que no ocurrió en la regresión logística multiclase.

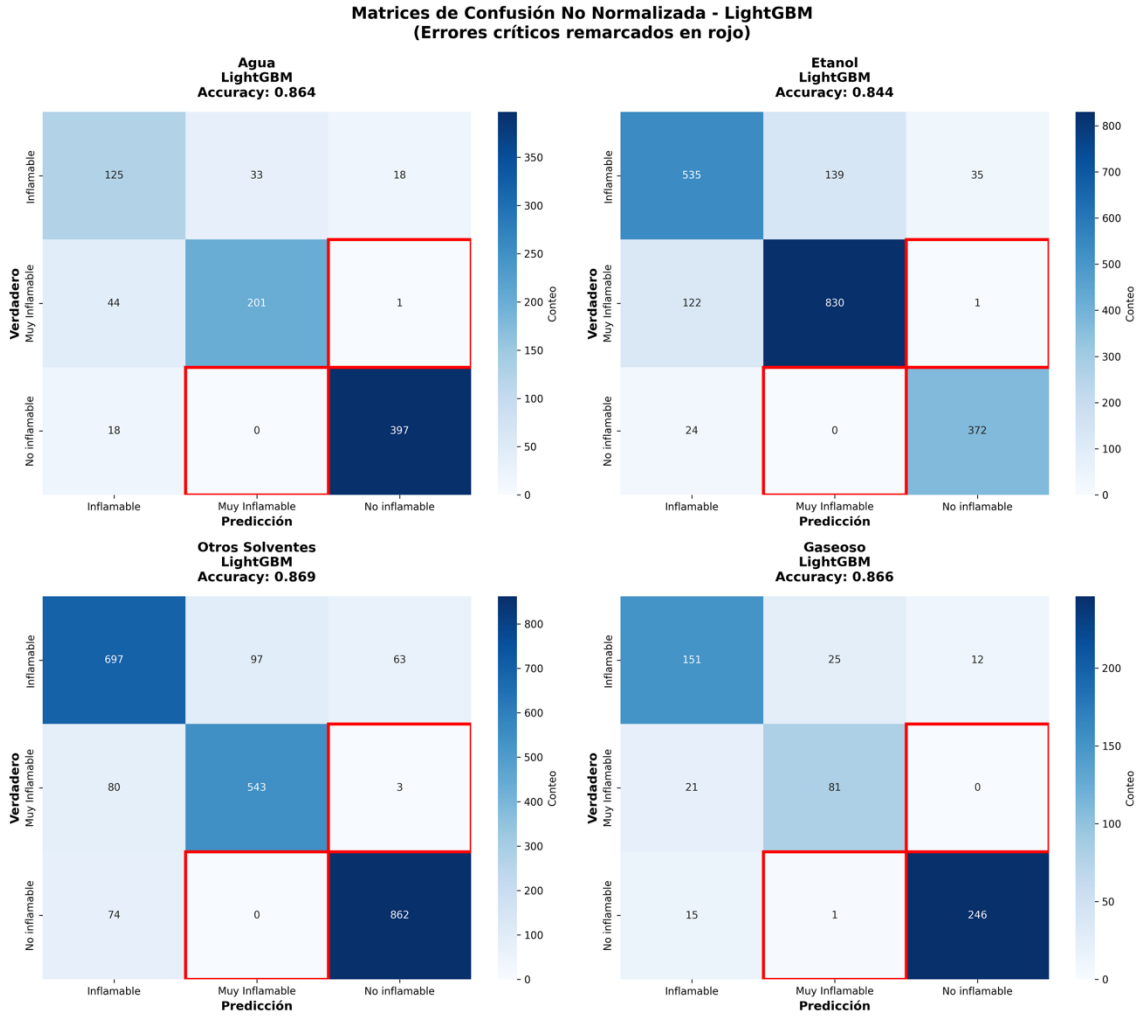

Figura N°20: Matrices de confusión para los grupos de fórmulas clasificados por el modelo LightGBM según su composición química.

En la Figura N°21 podemos comprobar que las fórmulas acuosas presentaron el peor F1-score de los cuatro grupos, siendo inferior al valor correspondiente a las soluciones acuosas modeladas mediante regresión logística multiclase. Para el caso de las fórmulas con solventes de bajo flash point se obtuvo el mejor F1-score de los modelos LightGBM.

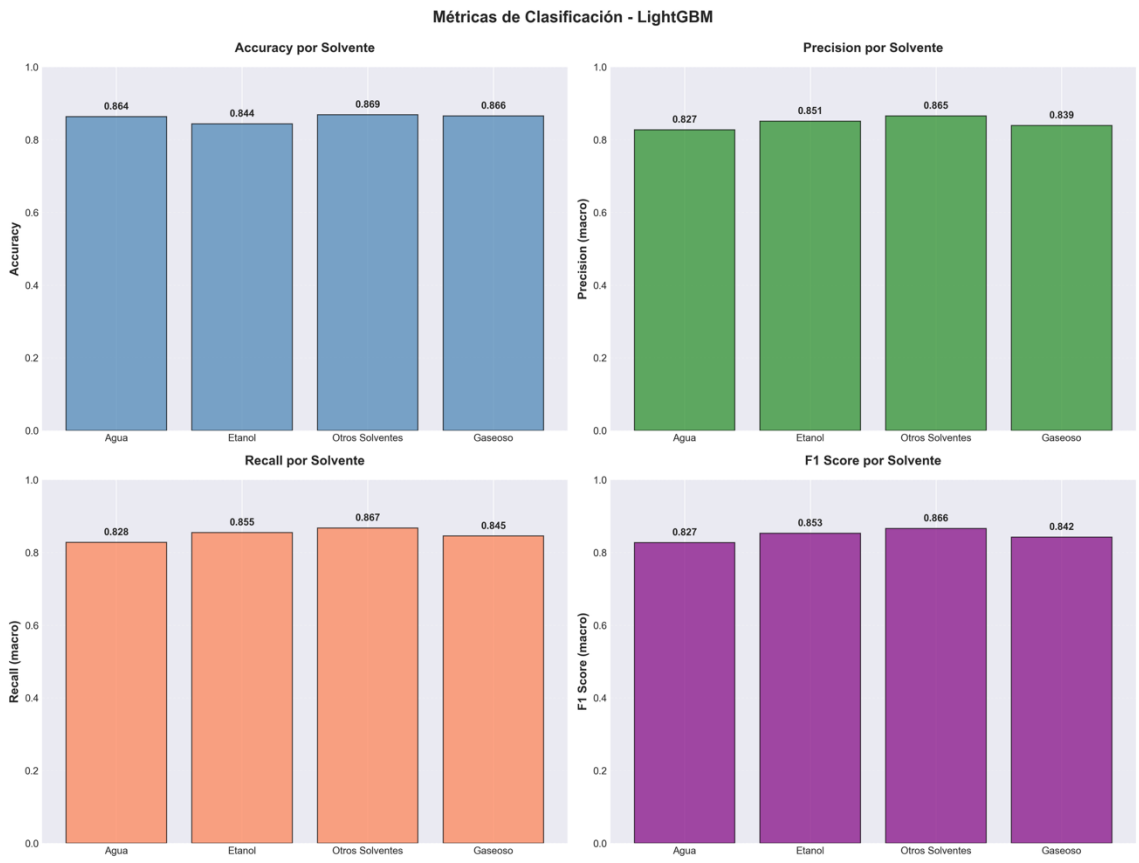

Figura N°21: Métricas de rendimiento del modelo LightGBM.

En la Tabla N°6 se comprueba que persiste el patrón relacionado con el menor rendimiento de las clases inflamables para las fórmulas acuosas, etanólicas, y que contienen solventes orgánicos de bajo flash point. Sin embargo, esto no se replica en el caso de las fórmulas con comportamiento gaseoso, obteniendo un menor rendimiento en la clase muy inflamable.

En todos los grupos las mejores métricas corresponden a las fórmulas no inflamables, lo cual debiera explicarse por su mayor presencia en los datasets, exceptuando el de fórmulas etanólicas, en el cual predomina la clase muy inflamable.

Grupo fórmula	Accuracy	Precision	Recall	F1-score
Acuoso muy inflamable	0,817	0,859	0,817	0,838
Acuoso inflamable	0,710	0,668	0,710	0,689
Acuoso no inflamable	0,957	0,954	0,957	0,955
Etanólico muy inflamable	0,871	0,857	0,871	0,864
Etanólico inflamable	0,755	0,786	0,755	0,770
Etanólico no inflamable	0,939	0,912	0,939	0,925
Gaseoso muy inflamable	0,794	0,757	0,794	0,775
Gaseoso inflamable	0,803	0,807	0,803	0,805
Gaseoso no inflamable	0,939	0,953	0,939	0,946
Otros solventes muy inflamable	0,867	0,848	0,867	0,858
Otros solventes inflamable	0,813	0,819	0,813	0,816
Otros solventes no inflamable	0,921	0,929	0,921	0,925

Tabla N°6: Métricas de clasificación LightGBM.

CatBoost con predicción conformal al 90% de confianza

Los modelos CatBoost fueron calibrados mediante predicción conformal, considerando un 90% de confianza. En la tabla N°7 se comparan las coberturas empíricas de los cuatro grupos de fórmulas, comprobando que no se logró el valor esperado para las soluciones etanólicas. En el caso de las fórmulas con ingredientes comportamiento de tipo gaseoso y correspondientes al grupo de solventes con bajo flash point se confirma que la cobertura empírica mejoró respecto al modelo LightGBM. Para las fórmulas acuosas disminuyó la cobertura empírica en comparación al modelo LightGBM. Sin embargo esta cobertura sigue siendo superior a 0,90. En resumen, los resultados generales obtenidos hasta este punto son más confiables en comparación a los resultados logrados por los modelos LightGBM.

Tipo de fórmula	Cobertura empírica
Acuosa	0,922
Etanólica	0,894
Comportamiento tipo gaseoso	0,909
Otros solventes	0,910

Tabla N°7: Cobertura empírica para CatBoost. El valor esperado para cada grupo es igual a 0,90.

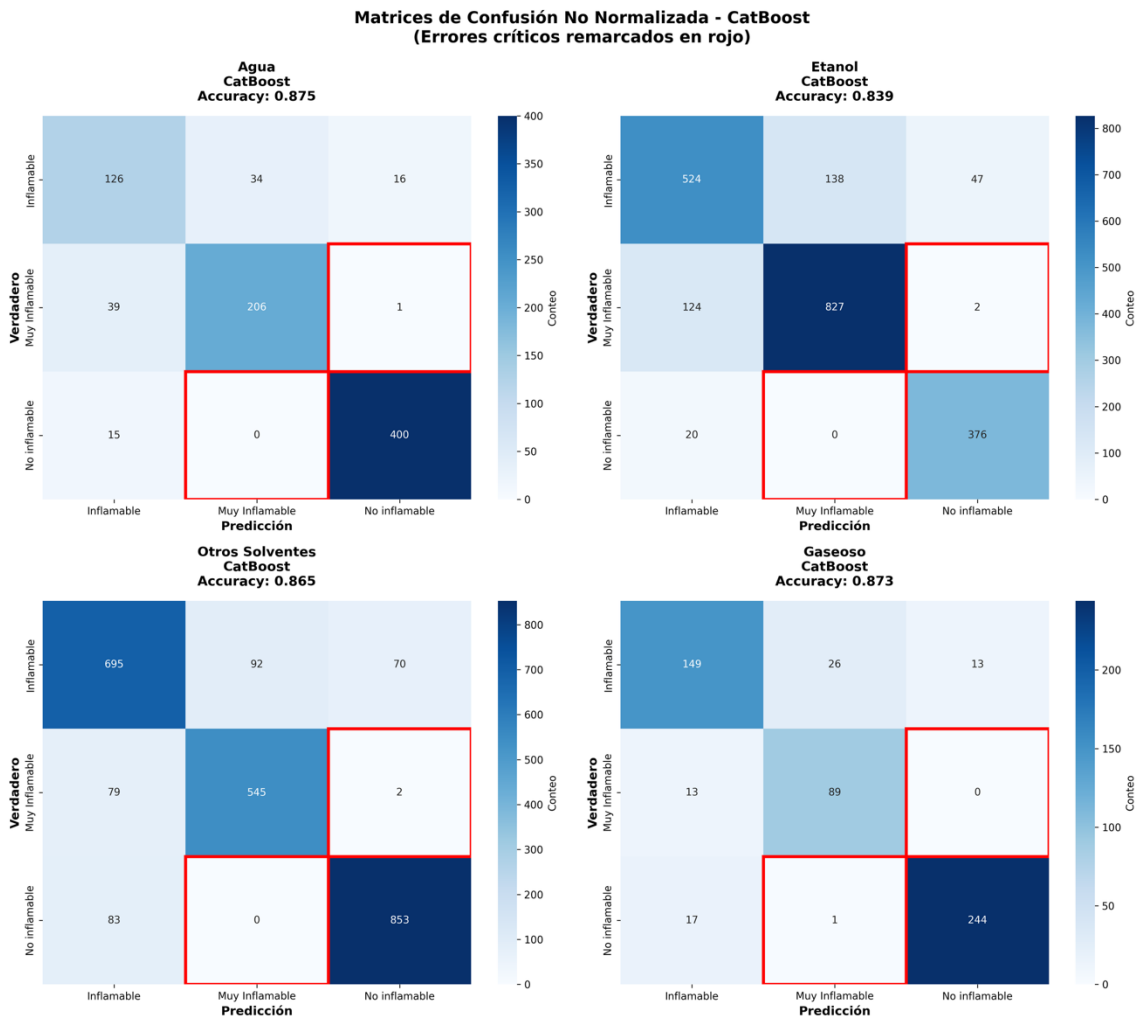

Figura N°22: Matrices de confusión para los grupos de fórmulas clasificados por el modelo CatBoost según su composición química.

En las matrices de confusión se comprueba que una fórmula acuosa, dos fórmulas etanólicas, y dos fórmulas que contienen otros solventes orgánicos fueron clasificadas erróneamente como no inflamables, cuando se trataba de fórmulas muy inflamables.

El caso contrario ocurrió con una fórmula que contiene ingredientes con comportamiento de tipo gaseoso.

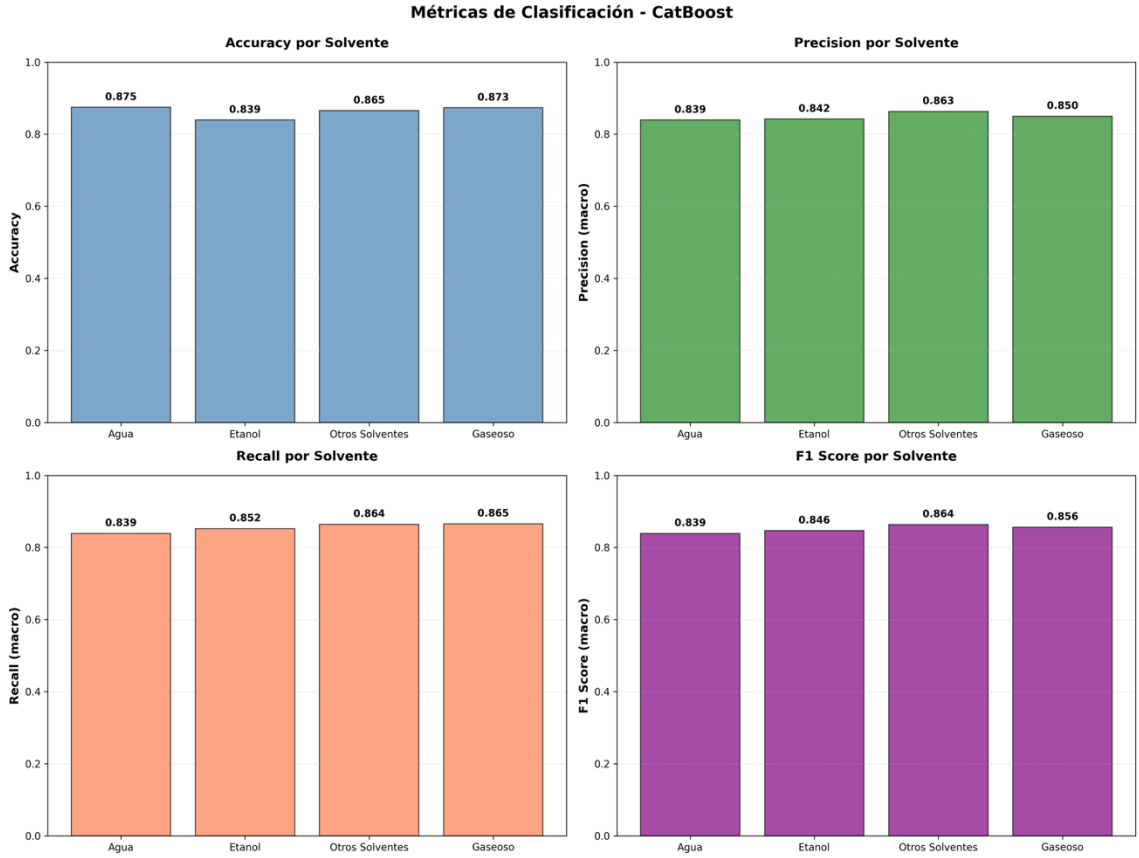

Figura N°23: Métricas de rendimiento del modelo CatBoost.

El mejor rendimiento general fue observado en el caso de las fórmulas que contienen solventes con bajo punto de inflamación, mientras que el valor más bajo de F1-score corresponde nuevamente a las fórmulas acuosas. Sin embargo, el rendimiento mejoró respecto a lo logrado por los modelos LightGBM y regresión logística multiclase.

En la Tabla N°8 se confirma que las fórmulas inflamables resultan ser más difíciles de clasificar correctamente en comparación a las fórmulas muy inflamables y no inflamables.

Grupo fórmula	Accuracy	Precision	Recall	F1-score
Acuoso muy inflamable	0,837	0,858	0,837	0,848
Acuoso inflamable	0,716	0,700	0,716	0,708
Acuoso no inflamable	0,964	0,959	0,964	0,962
Etanólico muy inflamable	0,868	0,857	0,868	0,862
Etanólico inflamable	0,739	0,784	0,739	0,761
Etanólico no inflamable	0,949	0,885	0,949	0,916
Gaseoso muy inflamable	0,873	0,767	0,873	0,817
Gaseoso inflamable	0,793	0,832	0,793	0,812
Gaseoso no inflamable	0,931	0,949	0,931	0,940
Otros solventes muy inflamable	0,871	0,856	0,871	0,863
Otros solventes inflamable	0,811	0,811	0,811	0,811
Otros solventes no inflamable	0,911	0,922	0,911	0,917

Tabla N°8: Métricas clasificación CatBoost.

Análisis de errores críticos de los modelos de clasificación multiclase

Error de clasificación	Comportamiento gaseoso	Solución etanólica	Solución acuosa	Otros solventes
Muy inflamable como no inflamable	No detectado	Tres fórmulas	Dos fórmulas	Cinco fórmulas
No inflamable como muy inflamable	No detectado	No detectado	No detectado	No detectado

Tabla N°9: Errores críticos regresión logística multiclase.

En el caso de los errores de las soluciones etanólicas, no se observa una tendencia relacionada con el porcentaje de solvente. La fórmulas detectadas contienen 0,12%, 20%, y un 45% de etanol.

Las soluciones acuosas clasificadas incorrectamente contienen un 25,92% y un 36% de agua respectivamente.

Respecto a las fórmulas que contienen otros solventes orgánicos de bajo flash point, los porcentajes totales de solvente son iguales o inferiores a un 5%. Una de ellas contiene un 1%, la segunda contiene un 0,20%, la tercera contiene un 5%, la cuarta contiene un 0,01%, y la quinta contiene un 0,27% de solvente.

Para las fórmulas con comportamiento gaseoso no se detectaron errores críticos de clasificación.

Error de clasificación	Comportamiento gaseoso	Solución etanólica	Solución acuosa	Otros solventes
Muy inflamable como no inflamable	No detectado	No detectado	Una fórmula	Tres fórmulas
No inflamable como muy inflamable	Una fórmula	No detectado	No detectado	No detectado

Tabla N°10: Errores críticos Random Forest.

Error de clasificación	Comportamiento gaseoso	Solución etanólica	Solución acuosa	Otros solventes
Muy inflamable como no inflamable	No detectado	Una fórmula	Una fórmula	Tres fórmulas
No inflamable como muy inflamable	Una fórmula	No detectado	No detectado	No detectado

Tabla N°11: Errores críticos LightGBM.

Se observó que ocurrieron casi los mismos errores de clasificación para los modelos Random Forest y LightGBM.

En el caso de los errores de las soluciones etanólicas, la fórmula detectada contiene 45% de etanol. Este error no fue detectado en el modelo Random Forest.

La fórmula acuosa clasificada incorrectamente como no inflamable contiene un 25,92% de agua.

Respecto a las fórmulas que contienen otros solventes de bajo flash point, los porcentajes totales de solvente son inferiores al 0,5%. La primera fórmula contiene un 0,20%, la segunda fórmula contiene un 0,43%, y la tercera contiene un 0,01% de solvente orgánico.

Para las fórmulas con comportamiento gaseoso se observó que una fórmula, la cual contiene un 4,33% de ingredientes con dicho comportamiento, fue clasificada como muy inflamable, siendo realmente una fórmula no inflamable.

Error de clasificación	Comportamiento gaseoso	Solución etanólica	Solución acuosa	Otros solventes
Muy inflamable como no inflamable	No detectado	Dos fórmulas	Una fórmula	Dos fórmulas
No inflamable como muy inflamable	Una fórmula	No detectado	No detectado	No detectado

Tabla N°12: Errores críticos CatBoost.

Los errores son similares a los del modelo LightGBM, aunque en el caso de CatBoost aparece nuevamente una fórmula etanólica que fue observada entre los errores del modelo de regresión lógistica multiclase. Además, sólo aparecen dos fórmulas en el caso de las fórmulas que contienen otros solventes.

Análisis SHAP modelos LightGBM y CatBoost

Análisis SHAP modelo LightGBM

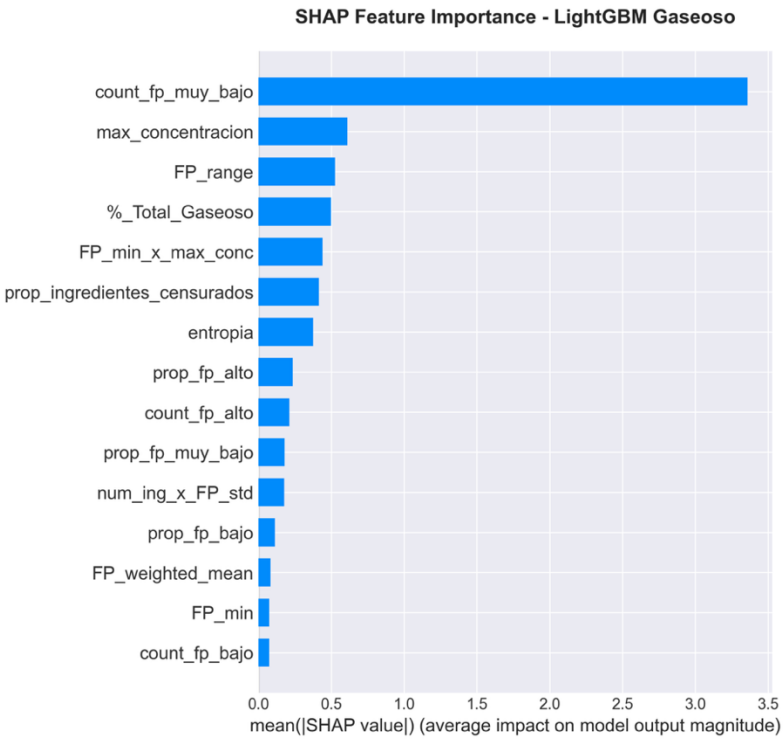

Figura N°24: Características más importantes del modelo LightGBM para el caso del dataset de fórmulas con comportamiento de tipo gaseoso.

En el caso del dataset de fórmulas que contienen ingredientes con comportamiento de tipo gaseoso, las variables con mayor impacto para la predicción de clases de inflamabilidad son la cantidad de ingredientes muy inflamables, la concentración del ingrediente dominante, el rango de flashpoint de una fórmula, el porcentaje de ingredientes con comportamiento de tipo gaseoso, y el flash point mínimo multiplicado por la concentración máxima.

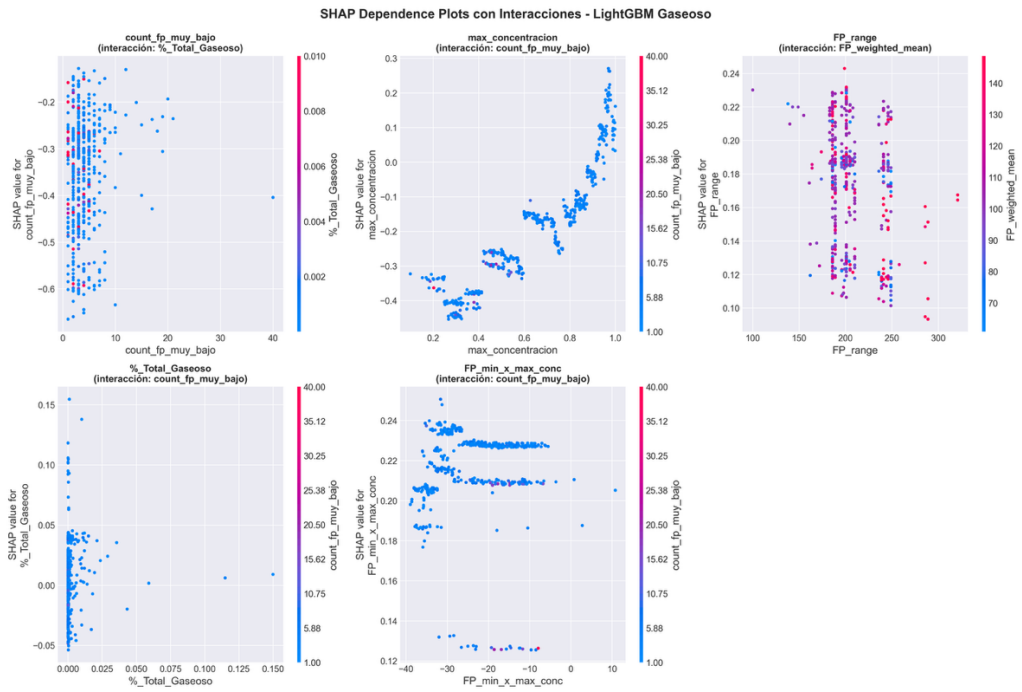

Figura N°25: Gráfico de dependencia del modelo LightGBM para el caso del dataset de fórmulas con comportamiento de tipo gaseoso.

La mayoría de los puntos de los valores SHAP se concentra verticalmente en valores iguales o menores a diez ingredientes muy inflamables en el gráfico que analiza la interacción con el porcentaje de solvente con comportamiento de tipo gaseoso.

En el gráfico de dependencia de los valores SHAP de la concentración del ingrediente dominante y su interacción con la cantidad de ingredientes muy inflamables, la tendencia es similar a la de una función cuadrática.

Para el caso del rango de flash point, los valores SHAP se concentran verticalmente entre 200 y 250 en el eje de las abscisas. Se confirma una clara interacción con el promedio ponderado del flash point.

Los valores SHAP se concentran verticalmente en porcentajes muy bajos de porcentaje total de ingredientes con comportamiento gaseoso. Existe una baja variabilidad en la interacción con la cantidad de ingredientes muy inflamables.

La variable flash point mínimo multiplicado por la concentración máxima muestra valores SHAP distribuidos verticalmente entre -40 y -30 respecto al eje de las abscisas, mientras que entre valores superiores a -30 y valores cercanos a cero se distribuyen horizontalmente.

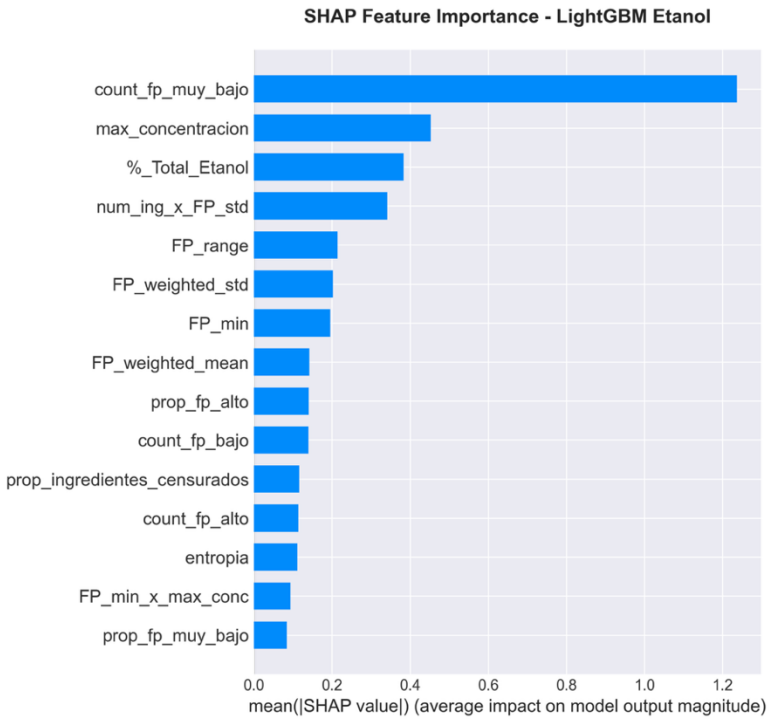

Figura N°26: Características más importantes del modelo LightGBM para el caso del dataset de fórmulas etanólicas.

En el caso del dataset de fórmulas etanólicas, las variables que impactan en mayor medida los resultados de clasificación del modelo LightGBM son la cantidad de

ingredientes muy inflamables, la concentración del ingrediente dominante, el porcentaje total de etanol de una fórmula, el número de ingredientes multiplicado por la desviación estándar del flash point, y el rango de flash point.

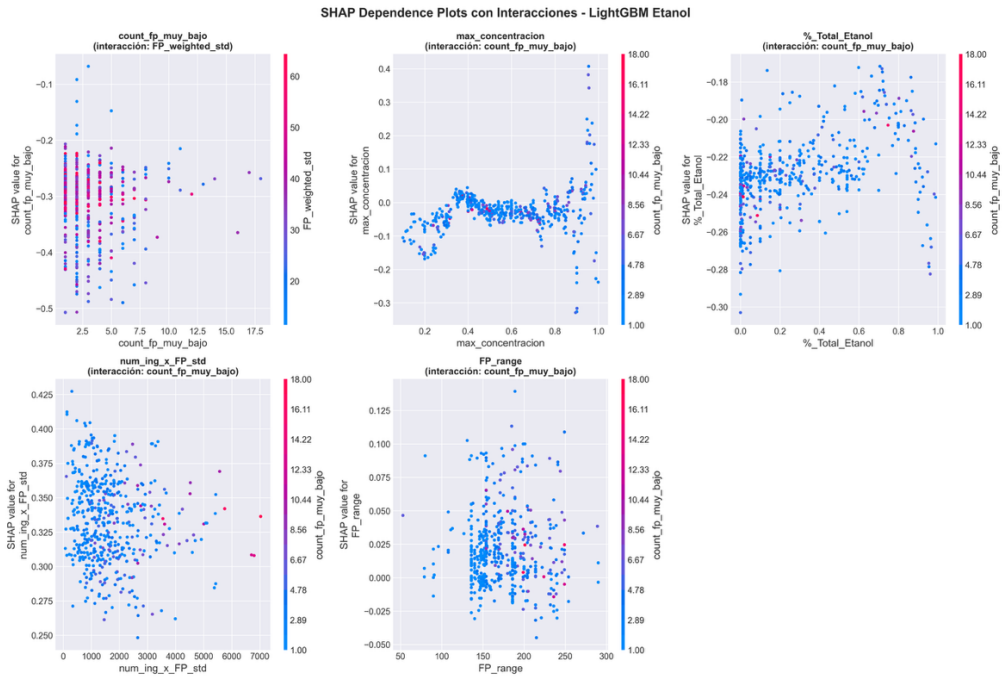

Figura N°27: Gráfico de dependencia del modelo LightGBM para el caso del dataset de fórmulas etanólicas.

En el gráfico de la variable cantidad de ingredientes muy inflamables los valores SHAP se distribuyen verticalmente, mostrando una tendencia a interactuar con valores medianos y altos de promedio ponderado del flash point.

La tendencia de los valores SHAP para la concentración del ingrediente dominante cambia de acuerdo al valor del eje de abscisas. La interacción suele ser con valores iguales o inferiores a diez ingredientes muy inflamables por fórmula.

En el caso de los valores SHAP del porcentaje total de etanol los puntos están bastante dispersos, observándose una tendencia a distribuirse verticalmente en el gráfico. El tipo de interacción con la cantidad de ingredientes muy inflamables es similar a la del gráfico de valores SHAP analizado anteriormente.

En el cuarto gráfico, correspondiente a la variable número de ingredientes multiplicado por la desviación estándar del flash point, se observa una nube de puntos, cuya mayor concentración se encuentra entre cero y 3000 respecto al eje de las abscisas. Un fenómeno similar ocurre en el gráfico de los valores SHAP del rango de flash point, cuyos resultados se concentran entre valores iguales a 130 y 250 en el eje de abscisas.

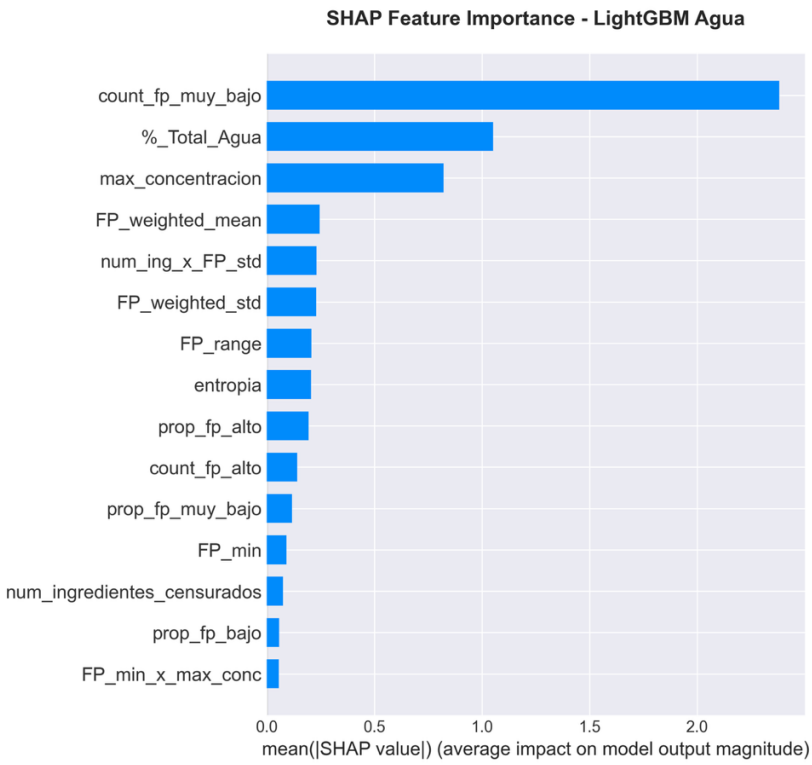

Figura N°28: Características más importantes del modelo LightGBM para el caso del dataset de fórmulas acuosas.

Al analizar las variables que influyen en la clasificación de las fórmulas acuosas, mostrado en la Figura N°28, se verifica que las que tienen mayor impacto son la cantidad de ingredientes muy inflamables, el porcentaje total de agua de una fórmula, la concentración del ingrediente dominante, el promedio ponderado del flash point, y el número de ingredientes multiplicado por la desviación estándar del flash point.

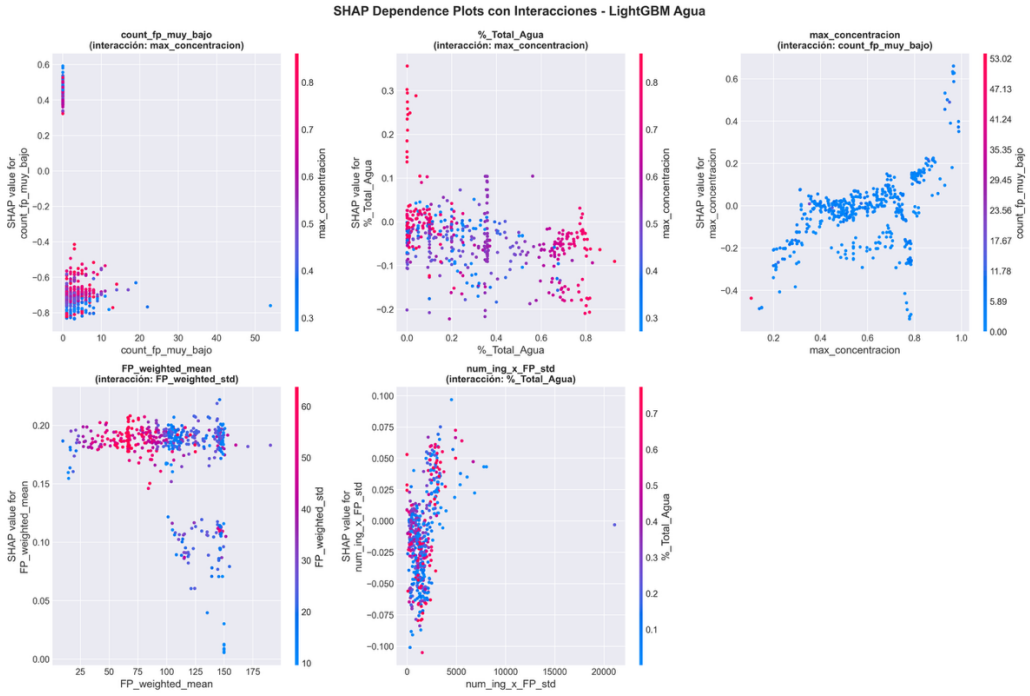

Figura N°29: Gráfico de dependencia del modelo LightGBM para el caso del dataset de fórmulas acuosas.

Los valores SHAP de la cantidad de ingredientes muy inflamables se concentra aproximadamente entre -0,8 y -0,4. La excepción ocurre con algunos resultados que se encuentran en un rango entre 0,3 y 0,6. Se observa una clara interacción con la variable concentración del ingrediente dominante.

En el caso del porcentaje total de agua los valores SHAP se distribuyen entre cero y 0,8 de forma mayoritaria en el eje de abscisas. La mayoría de las mediciones se encuentran entre -0,2 y 0,1. Se detecta una clara interacción con la concentración del ingrediente dominante de una fórmula.

En el gráfico de valores SHAP de la concentración del ingrediente dominante la tendencia depende del rango numérico del eje de abscisas. La tendencia es relativamente lineal con pendiente positiva.

Para los valores SHAP del promedio ponderado del flash point se comprueba que la mayoría de las cifras se concentra entre 0,15 y 0,20. En el eje de abscisas el rango se encuentra entre 20 y 150. Existe una clara dependencia con la desviación estándar ponderada por concentración.

El quinto gráfico muestra que los valores SHAP se concentran entre 0 y 5000 respecto a la variable número de ingredientes multiplicado por la desviación estándar del flash point. En el caso del eje de las ordenadas la distribución se encuentra entre -0,075 y 0,075.

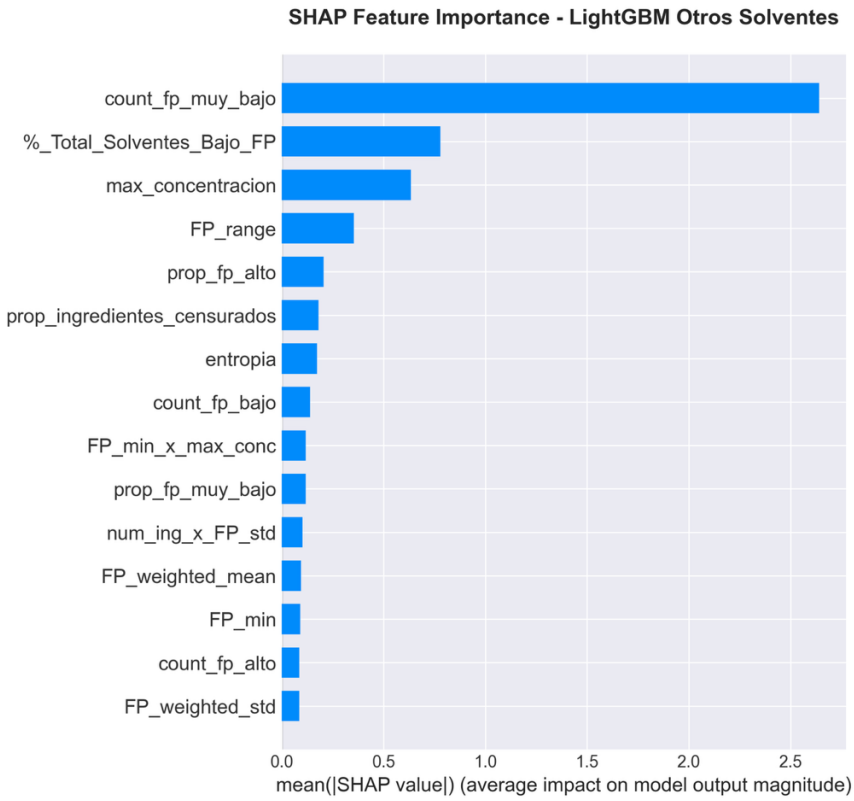

Figura N°30: Características más importantes del modelo LightGBM para el caso del dataset de fórmulas que contienen otros solventes.

Para el dataset de fórmulas que contienen solventes con bajo flash point, el análisis SHAP indica que las variables más relevantes para el modelo LightGBM son la cantidad de ingredientes muy inflamables, el porcentaje total de solvente de una fórmula, la concentración del ingrediente dominante, el rango de flash point, y la proporción de masa total de ingredientes con flash point mayor a 60°C.

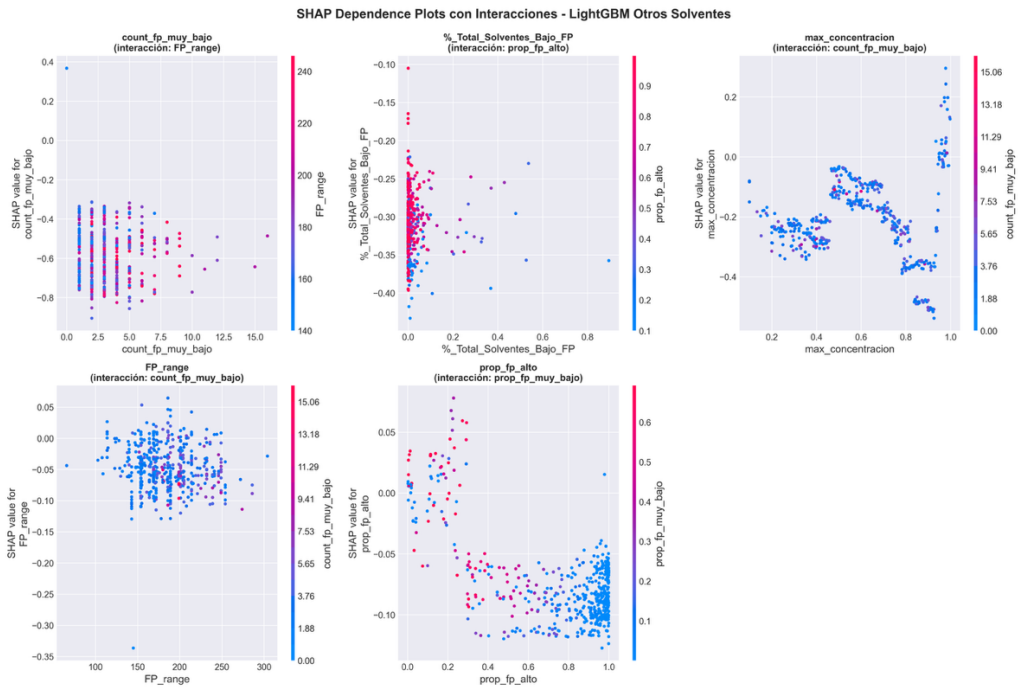

Figura N°31: Gráfico de dependencia del modelo LightGBM para el caso del dataset de fórmulas con otros solventes.

Los valores SHAP para la cantidad de ingredientes muy inflamables se distribuyen mayoritariamente entre -0,8 y -0,1. Los puntos muestran interacciones con la variable rango de flash point.

En el caso de los valores SHAP del porcentaje total de solvente, ellos se concentran en el rango entre 0 y 0,1 respecto al eje de abscisas. Al considerar el eje de las ordenadas como marco de referencia, las mediciones se distribuyen mayoritariamente entre -0,40 y -0,20.

Respecto a los valores SHAP de la concentración del ingrediente dominante, se detectan distintos comportamientos dependientes de dicha concentración. Sin embargo,

en un rango entre cero y aproximadamente 0,90 los valores SHAP se distribuyen en un rango negativo.

Los valores SHAP del rango de flash point ellos se distribuyen mayoritariamente entre -0,13 y 0,05. Respecto al eje de abscisas los puntos se concentran entre 100 y 250.

Al analizar el gráfico de valores SHAP de la proporción de masa total de ingredientes no inflamables se observan dos grupos. El primero se distribuye verticalmente entre 0 y 0,2, mientras que el segundo grupo se concentra entre 0,3 y 1, agrupándose entre los valores SHAP -0,15 y -0,05.

Análisis SHAP modelo CatBoost

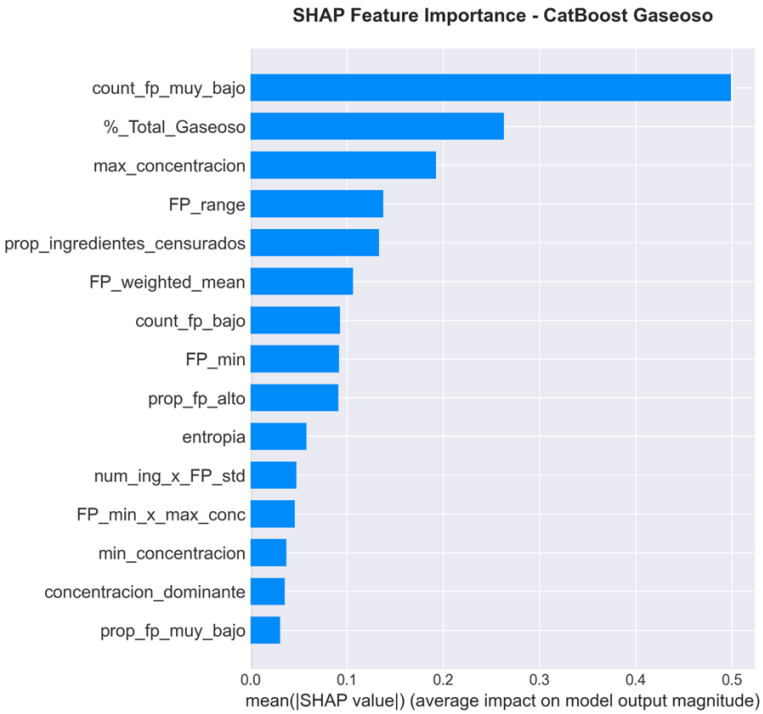

Figura N°32: Características más importantes del modelo CatBoost para el caso del dataset de fórmulas con comportamiento de tipo gaseoso.

Podemos verificar que las variables más importantes para el dataset de fórmulas con comportamiento gaseoso son la cantidad de ingredientes muy inflamables, el porcentaje de concentración de solventes gaseosos, la concentración del ingrediente dominante, el rango de flash point de una fórmula, y la proporción de ingredientes no inflamables cuyo flash point fue imputado dentro del total de ingredientes de una fórmula.

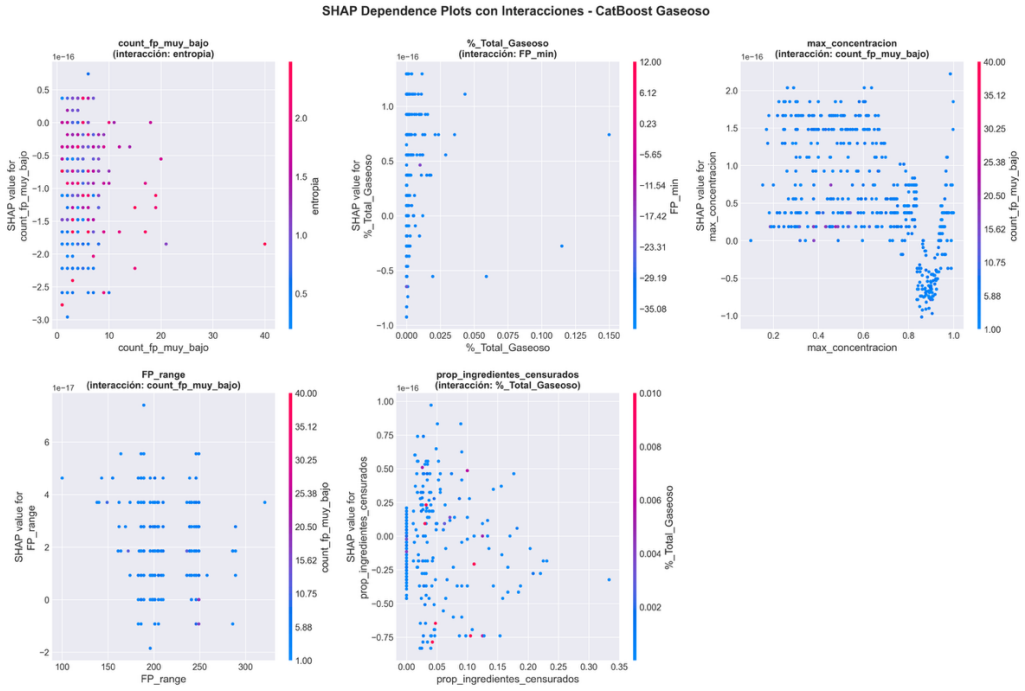

Figura N°33: Gráfico de dependencia del modelo CatBoost para el caso del dataset de fórmulas con componentes con comportamiento de tipo gaseoso.

Observamos que los valores SHAP de la cantidad de ingredientes muy inflamables se concentran en un rango entre 0 y 10 en el eje de las abscisas. Se comprueba que existe interacción con la variable entropía de Shannon.

Los valores SHAP del porcentaje de solvente con comportamiento de tipo gaseoso se concentra en porcentajes muy bajos. Su distribución es vertical.

En el caso de los valores SHAP de la concentración del ingrediente dominante los puntos se distribuyen entre cero y dos cuando la concentración es igual o inferior a 0,8.

En el gráfico de los valores SHAP del rango de flash point los puntos se distribuyen mayoritariamente entre 150 y 250 respecto al eje de las abscisas.

Respecto a los valores SHAP de la proporción de ingredientes no inflamables cuyo flash point fue imputado dentro del total de ingredientes de una fórmula, los puntos se distribuyen mayoritariamente entre cero y 0,15 respecto al eje de las abscisas.

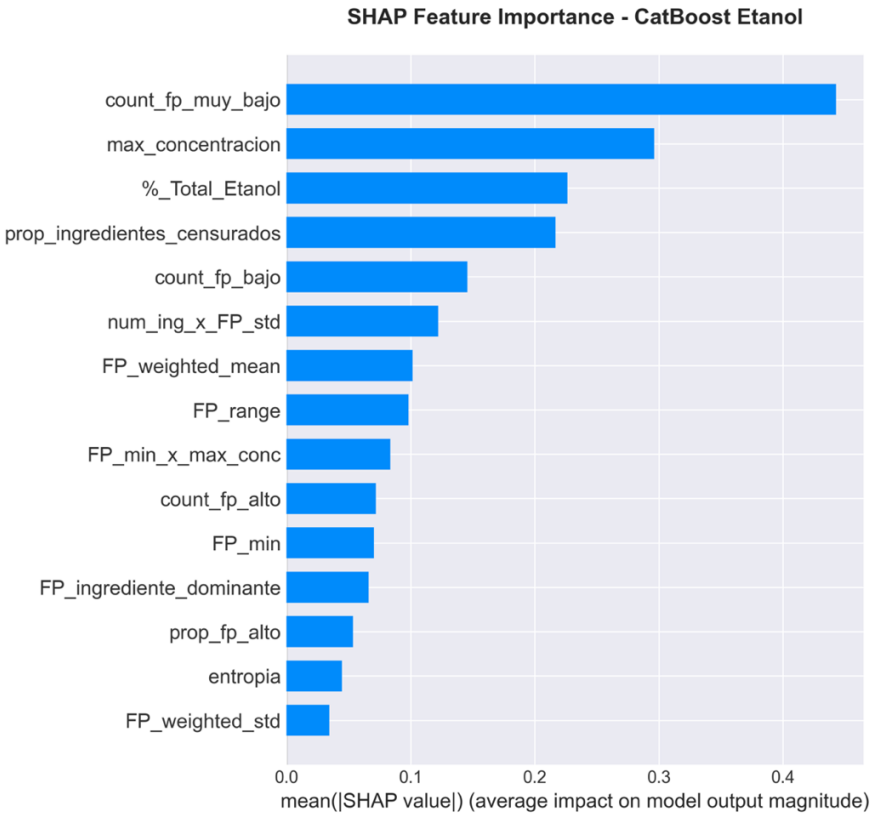

Figura N°34: Características más importantes del modelo CatBoost para el caso del dataset de fórmulas etanólicas.

El análisis SHAP determinó que las variables más importantes para el dataset de fórmulas etanólicas son la cantidad de ingredientes muy inflamables, la concentración del ingrediente mayoritario, el porcentaje total de etanol de una fórmula, la proporción de ingredientes no inflamables cuyo valor de flash point fue imputado dentro del total de ingredientes de una fórmula, y la cantidad de ingredientes inflamables de una fórmula.

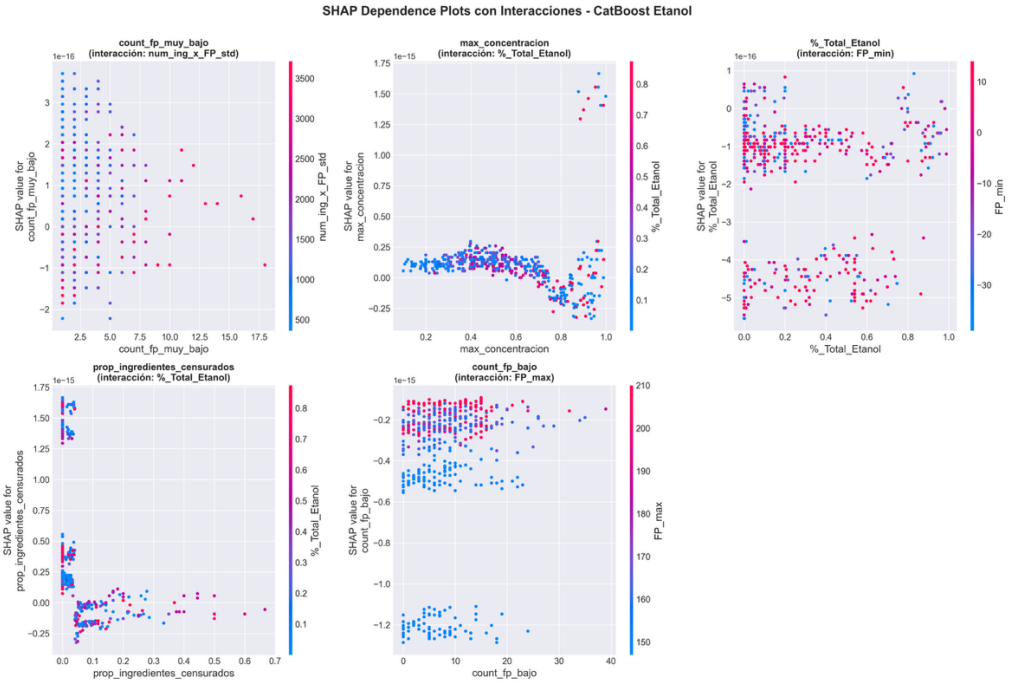

Figura N°35: Gráfico de dependencia del modelo CatBoost para el caso del dataset de fórmulas etanólicas.

Los valores SHAP de la cantidad de ingredientes muy inflamables presenta una distribución vertical y se concentran entre valores inferiores a 2,5 y 7,5 respecto al eje de las abscisas. Se verifica la interacción con el número de ingredientes multiplicado por la desviación estándar del flash point.

En el gráfico de la concentración del ingrediente dominante los valores SHAP se distribuyen entre cero y 0,25 mientras el valor de la concentración es igual o inferior a 0,70. Se comprueba que esta variable interactúa con el porcentaje total de etanol de una fórmula.

Respecto al gráfico de los valores SHAP del porcentaje total de etanol, se detecta un cluster concentrado entre cifras menores a -5 y -4, mientras que el segundo cluster se ubica entre -2 y 1. Esta variable interactúa con el flash point mínimo de una fórmula.

Para el caso del gráfico de los valores SHAP de la proporción de ingredientes no inflamables cuyo valor de flash point fue imputado dentro del total de ingredientes de una fórmula, se formaron varios clusters. Algunos clusters se localizan entre valores cercanos a cero, mientras que los clusters restantes se distribuyen en el rango de valores cercanos a cero y 0,5, observando valores escasos entre 0,6 y 0,7 respecto al eje de las abscisas. Esta variable interactúa con el porcentaje total de etanol.

El gráfico de valores SHAP de la cantidad de ingredientes inflamables un cluster se concentra entre -0,6 y -0,1, mientras que el otro cluster se localiza entre -1,3 y -1,1.

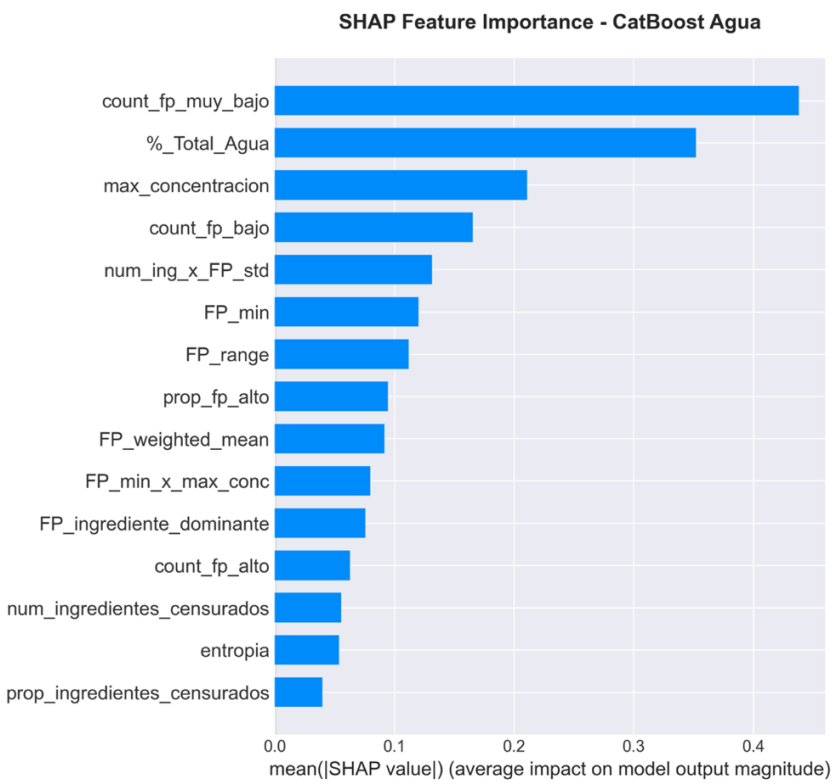

Figura N°36: Características más importantes del modelo CatBoost para el caso del dataset de fórmulas acuosas.

El análisis SHAP del dataset de fórmulas acuosas arrojó que las variables más importantes son la cantidad de ingredientes muy inflamables, el porcentaje total de agua de una fórmula, la concentración del ingrediente mayoritario, la cantidad de ingredientes inflamables, y el número de ingredientes multiplicado por la desviación estándar del flash point.

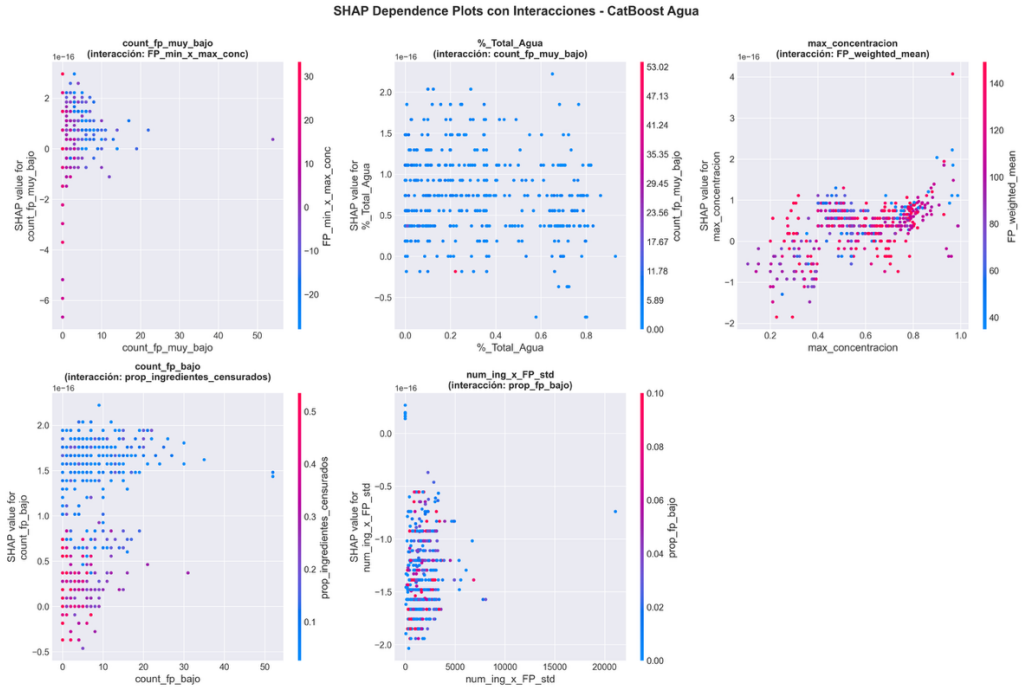

Figura N°37: Gráfico de dependencia del modelo CatBoost para el caso del dataset de fórmulas acuosas.

Los valores SHAP de la cantidad de ingredientes muy inflamables se concentran entre cero y diez respecto al eje de las abscisas. Las mediciones se concentran entre valores SHAP mayores a -2 y valores cercanos a 2. Esta variable interactúa con el flash point mínimo multiplicado por la concentración máxima.

En el gráfico del porcentaje total de agua los valores SHAP se concentran entre cero y dos. En el eje de las abscisas los puntos se concentran entre las cifras cero y 0,8. Esta variable interactúa con la cantidad de ingredientes muy inflamables de una fórmula.

Para el caso de los valores SHAP de la concentración del ingrediente mayoritario la mayoría de los puntos se concentra entre -1 y 1. Esta variable interactúa con el promedio ponderado del flash point.

Respecto a los valores SHAP de la cantidad de ingredientes inflamables, el rango numérico del eje de las abscisas se encuentra entre cero y treinta para la mayoría de las mediciones. Esta variable interactúa con la proporción de ingredientes no inflamables cuyo valor de flash point fue imputado dentro del total de ingredientes de una fórmula.

El gráfico de valores SHAP de la variable número de ingredientes multiplicado por la desviación estándar del flash point, las observaciones se concentran entre cero y cinco mil en el eje de las abscisas. Los valores SHAP mayoritariamente se encuentran en el rango entre -2 y -0,5. Esta variable interactúa con la proporción de masa total de ingredientes inflamables.

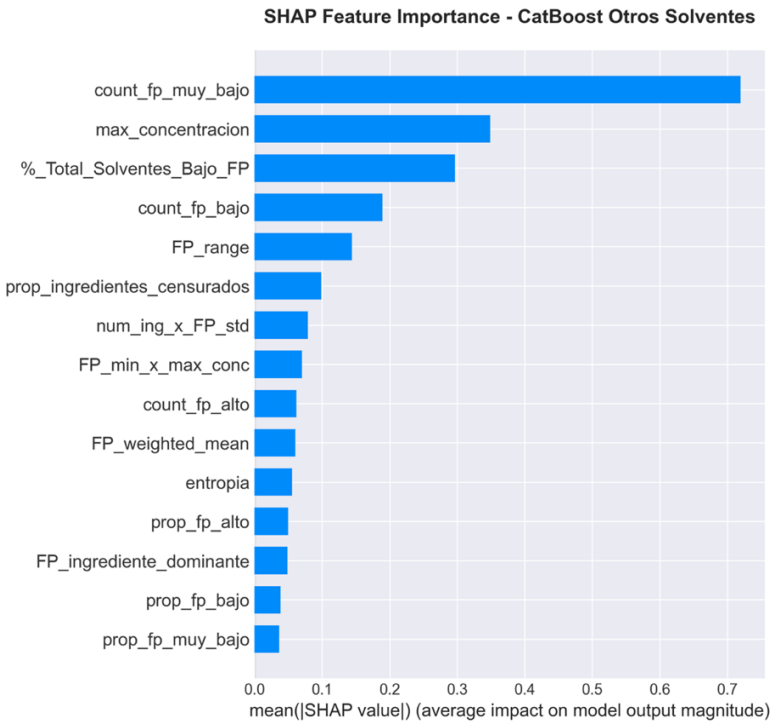

Figura N°38: Características más importantes del modelo CatBoost para el caso del dataset de fórmulas que contienen otros solventes.

El análisis SHAP determinó que las variables más importantes para el dataset de fórmulas que contienen solventes con bajo flash point son la cantidad de ingredientes muy inflamables, la concentración del ingrediente mayoritario, el porcentaje de solvente orgánico de una fórmula, la cantidad de ingredientes inflamables, y el rango de flash point de una fórmula.

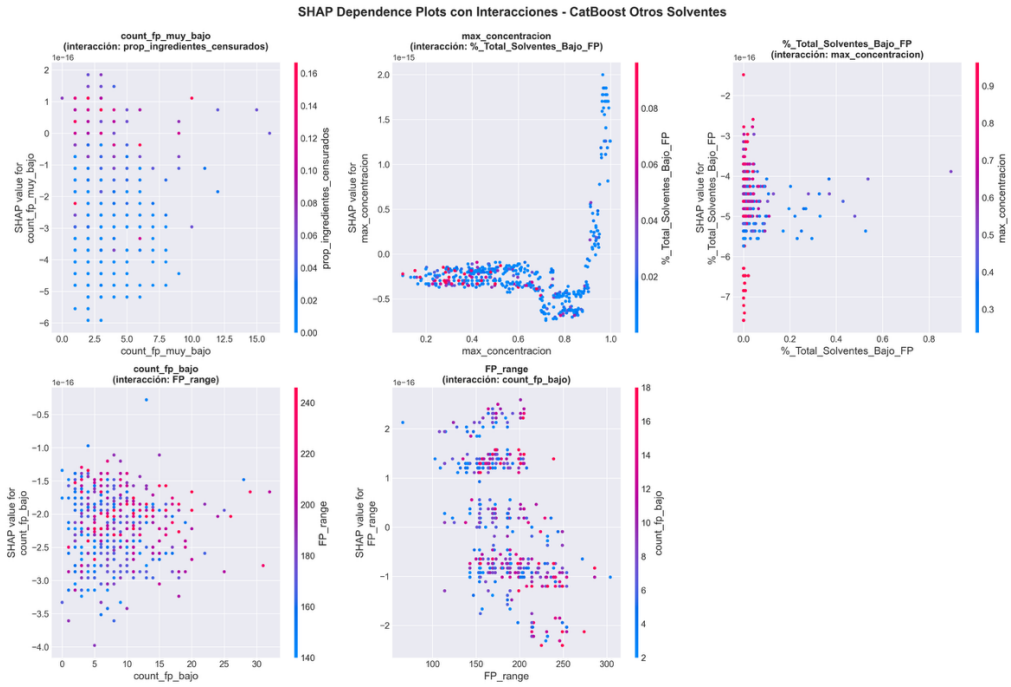

Figura N°39: Gráfico de dependencia del modelo CatBoost para el caso del dataset de fórmulas que contienen otros solventes.

En el gráfico de los valores SHAP de cantidad de ingredientes muy inflamables, la mayoría de las observaciones se distribuyen entre cero y 7,5 respecto al eje de las abscisas. Esta variable interactúa con la proporción de ingredientes no inflamables cuyo valor de flash point fue imputado dentro del total de ingredientes de una fórmula.

Para los valores SHAP de la concentración del ingrediente dominante, los valores SHAP se concentran entre -0,5 y cero cuando los valores de concentración se encuentran entre 0 y 0,7. Esta variable interactúa con el porcentaje total de solvente de bajo flash point.

En el caso de los valores SHAP del porcentaje total de solventes de bajo flash point, los puntos se concentran en un rango desde cero hasta 0,1 en el eje de las abscisas. Esta variable interactúa con la concentración del ingrediente dominante.

Respecto al gráfico de valores SHAP de la cantidad de ingredientes inflamables, se observa una nube de puntos cuyas observaciones se concentran entre cero y veinte respecto al eje de las abscisas. Esta variable interactúa con el rango de flash point de una fórmula.

El gráfico de valores SHAP del rango de flash point de una fórmula muestra que la gran mayoría de las observaciones se concentran entre cien y doscientos cincuenta respecto al eje de las abscisas. Esta variable interactúa con la cantidad de ingredientes inflamables de una fórmula.

6. Conclusiones

Durante el desarrollo de este trabajo fue posible comprobar que el rendimiento de los modelos de clasificación multiclase depende de las características químicas de las fórmulas analizadas. En general, se obtuvo mejores resultados al etiquetar las fórmulas no inflamables, mientras que se detectó que todos los modelos presentan problemas para clasificar correctamente las fórmulas inflamables. Además, tanto LightGBM como CatBoost tuvieron un menor rendimiento general al clasificar las fórmulas acuosas en comparación a los otros grupos químicos.

Se puede concluir que los modelos lograron clasificar adecuadamente la inflamabilidad de las fórmulas, independientemente de la cantidad de ingredientes contenidos por ellas. Esto constituye un avance respecto a los modelos enfocados en mezclas binarias y ternarias, debido a que posibilita la aplicación de este tipo de algoritmos para clasificar la inflamabilidad de otras mezclas complejas de acuerdo a sus propiedades fisicoquímicas.

Al comparar todos los modelos de clasificación multiclase se observó que el peor rendimiento total fue logrado por la regresión logística multiclase. Este resultado concuerda con el análisis de errores críticos, debido a que en total se observaron diez errores críticos en el caso de la regresión logística. En el caso de Random Forest este número disminuyó a cinco, y tanto para LightGBM como CatBoost se contabilizaron seis errores en total, considerando los cuatro grupos de fórmulas químicas en conjunto. Por lo

tanto, se puede concluir que los modelos basados en árboles de decisión interpretan de mejor manera el dataset de fórmulas de Cramer.

El análisis SHAP confirmó que la importancia de las características depende de las propiedades químicas de las fórmulas, como también del modelo de clasificación implementado. Esta diferencia podría explicar las diferencias de rendimiento observadas entre los modelos LightGBM y CatBoost, así como los errores críticos detectados para ambos algoritmos.

Se sugiere complementar a futuro este estudio con la implementación de modelos de regresión calibrados mediante predicción conformal para predecir el valor del punto de inflamación para los cuatro grupos de fórmulas químicas.

Respecto a las limitaciones del proyecto, estos modelos solamente consideraron datos de saborizantes y fragancias líquidas, por lo cual ellos no deben ser aplicados para clasificar la inflamabilidad de fórmulas sólidas, dado que sus propiedades fisicoquímicas difieren de las características de las fórmulas líquidas. Estos modelos de clasificación multiclase tampoco deben ser aplicados sin realizar ajustes previos -entrenamiento con datos adecuados, verificar las variables consideradas para el modelamiento del problema- a otros tipos de productos orgánicos volátiles, tales como combustibles.

Además, los datos presentan un desbalance de clases, lo cual probablemente explica el rendimiento inferior observado respecto a clasificación de las fórmulas inflamables.

Bibliografía

1.- Amirkhani F., Dashti A., Abedsoltan H., Mohammadi A. H., Chofreh A. G., Goni F. A., Klemeš J.J. (2022). Estimating flashpoints of fuels and chemical compounds using hybrid machine-learning techniques. *Fuel*, *323*, 124292. https://doi.org/10.1016/j.fuel.2022.124292

2.- Camesasca, M., Kaufman, M., Manas-Zloczower, I. (2006). Quantifying Fluid Mixing with the Shannon Entropy. *Macromolecular Theory and Simulations, 15*, 595-607. https://doi.org/10.1002/mats.200600037

3.- Cao W. et al. (2020). A novel method for predicting the flash points of binary mixtures from molecular structures. *Safety Science*, *126*, 104680. https://doi.org/10.1016/j.ssci.2020.104680

4.- Costa do Nascimiento, D., Prado de Omena Souza, M., de Oliveira Hentges, L., Macedo Dias, R., Barbosa Neto, A. M., Conceição da Costa, M. (2024). Mixture Flash Point Calculation: Recent Advances and a Closer Look at Biodiesel. ACS Chem. Health Saf., *31*(1), 22−43. https://doi.org/10.1021/acs.chas.3c00089

5.- Decreto 57 de 2021 [Ministerio de Salud]. Aprueba Reglamento de Clasificación, Etiquetado, y Notificación de Sustancias Químicas y Mezclas Peligrosas. 9 de febrero de 2021.

6.- Jovic, O., Mouras, R. (2024). Extreme Gradient Boosting Combined with Conformal Predictors for Informative Solubility Estimation, *Molecules*, *29*, 19. **https://doi.org/10.3390/molecules29010019**

7.- Manokhin, V. (2025). Applied conformal prediction: Practical uncertainity quantification for real-world ML. Leanpub. https://leanpub.com/advancedconformalprediction

8.- Mowbray, M., Vallerio, M., Perez-Galvan, C., Zhang, D., Del Rio Chanona, A., Navarro-Brull, F. J. (2022). Industrial data science– a review of machine learning applications for chemical and process industries. *React. Chem. Eng.*, *7*,1471. https://doi.org/10.1039/d1re00541c

9.- Pan , Y. y Juncheng Jiang, J. (2023). Flammability Characteristics Prediction Using QSPR Modeling en Q. Wang y C. Cai (Eds.), *Machine Learning in Chemical Safety and Health* (1ra ed., pp. 47-80). Wiley.

10.- Ponce-Bobadilla, A. V., Schmitt, V., Maier, C. S., Mensing, S., & Stodtmann, S. (2024). Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development. *Clinical and translational science*, *17*(11), e70056. https://doi.org/10.1111/cts.70056